MPS114 - An Introduction to Data Science

Dr Jill Johnson

2026-02-13

Contents

About these notes

These lecture notes are written for students on MPS114 Probability and Data Science. Topics covered include:

- data handling and exploratory analysis in R;
- a short introduction to machine learning;
- statistical inference: point estimation, interval estimation and hypothesis testing.

There are examples that you can try for yourself from Chapter 3 onwards. You will find these helpful for revision; exam questions will be in a similar style. If you are looking for an example on a particular topic, you may find the index of examples useful.

Chapter 1

Exploratory Data Analysis using R

In this chapter we study how to extract information from a data set using various plots and summary statistics. (We will do more formal statistical modelling and analysis in later chapters). We will study how to handle data sets in R and how to produce various plots.

1.1 Case study: what makes a country good at maths?

Every few years, the Organisation for Economic Co-operation and Development (OECD) conducts a survey, known as the Programme of International Student Assessment (PISA), to compare school systems across different countries. In the 2015 survey, 72 countries (including the UK) were compared, and about half a million 15-year-old children took tests in reading, mathematics and science. Two questions we may wish to consider are

1. how does the UK compare with other countries?
2. Are there factors that could explain why some countries do better than others?

The PISA data can be obtained from (http://pisadataexplorer.oecd.org/ide/idepisa/). To consider the second question, we will explore some of the data available from the World Bank

A spreadsheet `maths.csv` (which you can download here) has been compiled from these sources, and gives five variables for each country

1. `score`: the mean mathematics score in the 2015 PISA test;
2. `gdp`: the gross domestic product per capita (GDP divided by the estimated population size), measured in US$;
3. `gini`: the gini coefficient (as a percentage). This is an estimate of income inequality, with larger values indicating more income inequality;
4. `homework`: an estimate of the average number of hours per week spent on homework by 15 year-olds, from a survey in 2012;
5. `start.age`: the age (in years) in which children start school.

In the remainder of this chapter, we will see how an exploratory data analysis can be used to extract information from our data, to help answer the two questions above.

1.2 The "Tidyverse"

The Tidyverse is a collection of R packages designed for data science.

(Artwork by @allison_horst)

We will be using some of these packages (in particular, a package for data manipulation called `dplyr`, and a package for graphics called `ggplot2`) in this course. You will need to install it with the command

```
install.packages("tidyverse")
```

You only need to do this once, but every session, you will need to load it with the command

```
library(tidyverse)
```

1.3 Importing data into R: `csv` and `.xlsx` files

The data are in "comma separated variables" (csv) format, so we can use the command `read_csv` to get the data into R. (The file `maths.csv` will need to be in your working directory. You can change the working directory in RStudio by going to Session > Set Working Directory).

```
maths <- read_csv("maths.csv")
```

The data are now stored in R in an object called `maths`. All commands and names in R are case-sensitive: R won't recognise the name `Maths`.

In R, you can read in files directly from websites. If you are working through this chapter on your own, and want to try out all the commands, you will either need to download the file `maths.csv` first, or you can import it directly with the command

```
maths <- read_csv("https://oakleyj.github.io/exampledata/maths.csv")
```

1.3.1 Importing Excel `.xlsx` files

If your data is an Excel spreadsheet in `.xlsx` format, you can either save it in Excel as a `.csv` file, or you can use the `read_xlsx()` command (for which you will need to load the `readxl` package first). For example, to import a file `spreadsheet.xlsx`, you would use the command

```
library(readxl)
mydata <- read_xlsx("spreadsheet.xlsx")
```

1.4 Data frames and tibbles in R

`maths` is a type of object known as a data frame, which is the main way of organising data in R. In fact, it's actually a special type of data frame known as a "tibble". We'll always be using tibbles rather than ordinary data frames here, but we won't worry too much about the difference.

The data are arranged in the data frame as they were in the .csv file, with one row per country, and one column per variable. Typing `maths` (and pressing return) in the R console will display the first ten rows only, and as many columns as will fit in the window.

```
maths
```

```
## # A tibble: 70 x 7
##    country          continent     score   gdp  gini homework start.age
##    <chr>            <chr>         <dbl> <dbl> <dbl>    <dbl>     <dbl>
##  1 Albania          Europe          413  4147  29        5.1          6
##  2 Algeria          Africa          360  3844  27.6     NA            6
##  3 Argentina        South America   409 12449  42.7      3.7          6
##  4 Australia        Oceanea         494 49928  34.7      6            5
##  5 Austria          Europe          497 44177  30.5      4.5          6
##  6 B-S-J-G (China)  Asia            531  8123  42.2     13.8          6
##  7 Belgium          Europe          507 41096  28.1      5.5          6
##  8 Brazil           South America   377  8650  51.3      3.3          6
##  9 Bulgaria         Europe          441  7351  37.4      5.6          7
## 10 Canada           North America   516 42158  34        5.5          6
## # i 60 more rows
```

Within the data frame, we see that the column names are `country`, `continent`, `score`, `gdp`, `gini`, `homework` and `start.age`, which we will use in various commands described shortly.

If we want to see all 70 rows, we can either use the command

```
print(maths, n = 70)
```

or, in RStudio, we can click on `maths` in the Environment window.

1.4.1 Ordering the rows by a variable with the `arrange()` command

Suppose we want to see which countries got the highest score: we want to arrange the rows in the dataframe `maths` in order according to the values in the column `score`. To do this we use the command

```
maths %>%
  arrange(score)
```

```
## # A tibble: 70 x 7
##    country            continent     score   gdp  gini homework start.age
##    <chr>              <chr>         <dbl> <dbl> <dbl>    <dbl>     <dbl>
##  1 Dominican Republic North America   328  6722  44.9     NA           6
##  2 Algeria            Africa          360  3844  27.6     NA           6
##  3 Tunisia            Africa          367  3689  35.8      3.5         6
##  4 Macedonia, FYR     Europe          371  5237  35.6     NA           6
##  5 Brazil             South America   377  8650  51.3      3.3         6
##  6 Jordan             Asia            380  4088  33.7      4.2         6
##  7 Indonesia          Asia            386  3570  39.5      4.9         7
##  8 Peru               South America   387  6046  44.3      5.5         6
##  9 Colombia           South America   390  5806  51.1      5.3         6
## 10 Lebanon            Asia            396  7914  31.8      3.3         6
## # i 60 more rows
```

This has arranged the rows in ascending order of `score`. To see them in descending order, we include the `desc()` command:

```
maths %>%
  arrange(desc(score))
```

```
## # A tibble: 70 x 7
##    country              continent     score   gdp  gini homework start.age
##    <chr>                <chr>         <dbl> <dbl> <dbl>    <dbl>     <dbl>
##  1 Singapore            Asia            564 52961  NA        9.4         6
##  2 Hong Kong SAR, China Asia            548 43681  NA        6           6
##  3 Macao SAR, China     Asia            544 73187  NA        5.9         6
##  4 Japan                Asia            532 38894  32.1      3.8         6
##  5 B-S-J-G (China)      Asia            531  8123  42.2     13.8         6
##  6 Korea, Rep.          Asia            524 27539  31.6      2.9         6
##  7 Switzerland          Europe          521 78813  32.5      4           7
##  8 Estonia              Europe          520 17575  34.6      6.9         7
##  9 Canada               North America   516 42158  34        5.5         6
## 10 Netherlands          Europe          512 45295  28.6      5.8         6
## # i 60 more rows
```

(Just by sorting the data, we can see something interesting: look at the `continent` variable for the highest scoring countries.)

1.4.2 Selecting rows with the `filter()` command

If we want to view a subset of the rows, we can use the `filter()` command. For example, if we want the rows in the data frame `maths` where `start.age` takes the value 5 (i.e. children start school at age 5), we can do

```
maths %>%
  filter(start.age == 5)
```

```
## # A tibble: 6 x 7
##   country            continent     score   gdp  gini homework start.age
##   <chr>              <chr>         <dbl> <dbl> <dbl>    <dbl>     <dbl>
```

```
## 1 Australia           Oceanea         494 49928  34.7      6           5
## 2 Ireland             Europe          504 61606  31.9      7.3         5
## 3 Malta               Europe          479 25058  NA       NA           5
## 4 New Zealand         Oceanea         495 39427  NA        4.2         5
## 5 Trinidad and Tobago North America   417 15377  40.3     NA           5
## 6 United Kingdom      Europe          492 39899  34.1      4.9         5
```

Note the double equals sign `==`. This is used to test whether the left and right hand sides are equal: each country is included if its corresponding `start.age` is equal to 5. The UK is included above, but we'll give an example of selecting it anyway:

```
maths %>%
  filter(country == "United Kingdom")
```

```
## # A tibble: 1 x 7
##   country        continent score   gdp  gini homework start.age
##   <chr>          <chr>     <dbl> <dbl> <dbl>    <dbl>     <dbl>
## 1 United Kingdom Europe      492 39899  34.1      4.9         5
```

1.4.3 Viewing and extracting data from a column

For larger data frames (with many columns), we may wish to view a subset only. For example, to select the `score` and `country` columns only from the `maths` data frame, we do

```
maths %>%
  select(score, country)
```

```
## # A tibble: 70 x 2
##    score country
##    <dbl> <chr>
##  1   413 Albania
##  2   360 Algeria
##  3   409 Argentina
##  4   494 Australia
##  5   497 Austria
##  6   531 B-S-J-G (China)
##  7   507 Belgium
##  8   377 Brazil
##  9   441 Bulgaria
## 10   516 Canada
## # i 60 more rows
```

If we want to extract the values from a column, we use the syntax `dataframe-name$column-name`. For example, to extract the column `score` from the data frame `maths`, we do

```
maths$score
```

```
##  [1] 413 360 409 494 497 531 507 377 441 516 423 390 400 464 437 492 511 328 520
## [20] 511 493 404 506 454 548 477 488 386 504 470 490 532 380 460 524 482 396 478
## [39] 486 544 371 446 479 408 420 418 512 495 502 387 504 492 402 444 494 564 475
## [58] 510 486 494 521 415 417 367 420 427 492 470 418 495
```

We could then, for example, calculate the average (mean) of all the scores:

```
mean(maths$score)
```

```
## [1] 460.9714
```

1.4.4 Creating new columns in a data frame with the `mutate()` command

About half the countries have a GDP per capita greater than $17000. If we try the command

```
maths$gdp > 17000
```

```
##  [1] FALSE FALSE FALSE  TRUE  TRUE FALSE  TRUE FALSE FALSE  TRUE FALSE FALSE
## [13] FALSE FALSE  TRUE  TRUE  TRUE FALSE  TRUE  TRUE  TRUE FALSE  TRUE  TRUE
## [25]  TRUE FALSE  TRUE FALSE  TRUE  TRUE  TRUE  TRUE FALSE FALSE  TRUE FALSE
## [37] FALSE FALSE  TRUE  TRUE FALSE FALSE  TRUE FALSE FALSE FALSE  TRUE  TRUE
## [49]  TRUE FALSE FALSE  TRUE  TRUE FALSE FALSE  TRUE FALSE  TRUE  TRUE  TRUE
## [61]  TRUE FALSE FALSE FALSE FALSE  TRUE  TRUE  TRUE FALSE FALSE
```

this creates a new vector, in which the ith element will be `TRUE` if the `gdp` value for country i is greater than 17000, and `FALSE` otherwise. We will add this vector to the data frame, under the column name `wealthiest`. The command to create the new column is

```
maths %>%
  mutate(wealthiest = maths$gdp > 17000)
```

but this doesn't store the result. To put the new column in the `maths` dataframe, we do

```
maths <- maths %>%
  mutate(wealthiest = maths$gdp > 17000)
```

You may now have too many columns to see in your console, so to check this has worked, we will do

```
maths %>%
  select(country, gdp, wealthiest)
```

```
## # A tibble: 70 x 3
##    country           gdp wealthiest
##    <chr>           <dbl> <lgl>
##  1 Albania          4147 FALSE
##  2 Algeria          3844 FALSE
##  3 Argentina       12449 FALSE
##  4 Australia       49928 TRUE
##  5 Austria         44177 TRUE
##  6 B-S-J-G (China)  8123 FALSE
##  7 Belgium         41096 TRUE
##  8 Brazil           8650 FALSE
##  9 Bulgaria         7351 FALSE
## 10 Canada          42158 TRUE
## # i 60 more rows
```

1.4.5 Chaining commands together with the pipe operator %>%

We've been making use of the 'pipe operator' %>%, which we will now discuss a little more. The pipe operator %>% takes whatever the output is from the left hand side, and uses it as the first argument in the function on the next line. In general,

```
myfunction(x, y)
```

is the same as

```
x %>%
  myfunction(y)
```

so, for example,

```
maths %>%
  filter(continent == "Europe")
```

is the same as `filter(maths, continent == "Europe")`.

The %>% syntax can be easier to read when we want to chain several functions together. Suppose we want to see the top 10 countries by score, within Europe. This will involve using both the `arrange()` and `filter()` commands. We can chain these together as follows

```
maths %>%
  filter(continent == "Europe") %>%
  arrange(desc(score))
```

```
## # A tibble: 39 x 8
##    country     continent score   gdp  gini homework start.age wealthiest
##    <chr>       <chr>     <dbl> <dbl> <dbl>    <dbl>     <dbl> <lgl>
##  1 Switzerland Europe      521 78813  32.5      4           7 TRUE
##  2 Estonia     Europe      520 17575  34.6      6.9         7 TRUE
##  3 Netherlands Europe      512 45295  28.6      5.8         6 TRUE
##  4 Denmark     Europe      511 53418  28.5      4.3         6 TRUE
##  5 Finland     Europe      511 43090  26.8      2.8         7 TRUE
##  6 Slovenia    Europe      510 21305  25.7      3.7         6 TRUE
##  7 Belgium     Europe      507 41096  28.1      5.5         6 TRUE
##  8 Germany     Europe      506 41936  31.4      4.7         6 TRUE
##  9 Ireland     Europe      504 61606  31.9      7.3         5 TRUE
## 10 Poland      Europe      504 12372  32.1      6.6         7 FALSE
## # i 29 more rows
```

We read this as, "Start with the `maths` data set, filter it to extract the European countries, then arrange in order of decreasing scores." The code above could be written as

```
arrange(filter(maths, continent == "Europe"), desc(score))
```

but a single (and potentially long) line of code such as this can be harder to read and understand.

1.5 Calculating summary statistics with the `summary()` command

We can now start to compare the UK's score with that in other countries. Recall that `maths$score` extracts the maths scores for the 70 countries. We can obtain various summary statistics for this variable with the command

```
summary(maths$score)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   328.0   417.2   477.5   461.0   500.8   564.0
```

We'll need a few definitions to interpret some of this output.

Definition 1.1 (Percentile/Quantile)**.** Given a set of (real-valued) observations, for any $\alpha \in [0, 1]$, the α quantile or 100α percentile is a value where (approximately) $100\alpha\%$ of the observations are below this value, and the remainder are above.

So, for example,

- the 25th percentile, or 0.25 quantile, is 417.2: approximately 25% of the countries have scores less than (or equal to) 417.2;
- the 75th percentile, or 0.75 quantile, is 500.8: approximately 75% of the countries have scores less than (or equal to) 500.8.

(There are different algorithms for obtaining the value of percentile/quantile, for example using linear interpolation, but we won't worry with the details here.)

Definition 1.2 (Median and quartiles)**.** The median is the 50th percentile/0.5 quantile. The lower quartile is the 25th percentile/0.25 quantile, and the upper quartile is the 75th percentile/0.75 quantile.

The output from the `summary()` command also tells us that

- the smallest observed score was 328;
- the median score was 477.5;
- the arithmetic mean (sum of all the scores, divided by 70) for the 70 countries was 461;
- the largest observed score was 564.

Definition 1.3 (The interquartile range)**.** The interquartile range is the difference between the 75th percentile (0.75 quantile) and 25th percentile (0.25 quantile), and is sometimes used to describe variation in data.

The interquartile range can be obtained in R with the command

```
IQR(maths$score)
```

```
## [1] 83.5
```

We've seen that the UK's score was 492. This ranks the UK outside the top 25%, but inside the top 50%. We can find the actual rank as follows:

```
sum(maths$score > 492)
```

```
## [1] 25
```

This tells us that the UK ranked 26th: 25 countries got higher scores.

1.5.1 Calculating individual summary statistics

Individually, we could have calculated these summaries as follows

```
min(maths$score)
```

```
## [1] 328
```

```
quantile(maths$score, 0.25)
```

```
##    25%
## 417.25
```

```
quantile(maths$score, 0.5)
```

```
##   50%
## 477.5
```

```
mean(maths$score)
```

```
## [1] 460.9714
```

```
quantile(maths$score, 0.75)
```

```
##    75%
## 500.75
```

```
max(maths$score)
```

```
## [1] 564
```

This may be more convenient if we just want a particular summary statistic, and/or want to store the result for use later on.

1.5.2 Calculating other quantiles/percentiles

If we wanted, say, the 90th percentile (0.9 quantile), we could do

```
quantile(maths$score, 0.9)
```

```
##   90%
## 520.1
```

so that, approximately, 90% of the countries has scores less than or equal to 520.1

1.5.3 Computing summaries per group

Suppose we want to know the mean score within particular groups, for example, continents. We can do this by chaining together the `group_by()` and `summarise()` commands.

```
maths %>%
  group_by(continent) %>%
  summarise(meanscore = mean(score))
```

```
## # A tibble: 6 x 2
##   continent     meanscore
##   <chr>             <dbl>
```

```
## 1 Africa            364.
## 2 Asia              471.
## 3 Europe            476.
## 4 North America     423.
## 5 Oceanea           494.
## 6 South America     401.
```

We read the command as, "Start with the `maths` data frame, organise into groups based on the `continent` column, then create a new variable called `meanscore`, which is the mean of the `score` variable within each group."

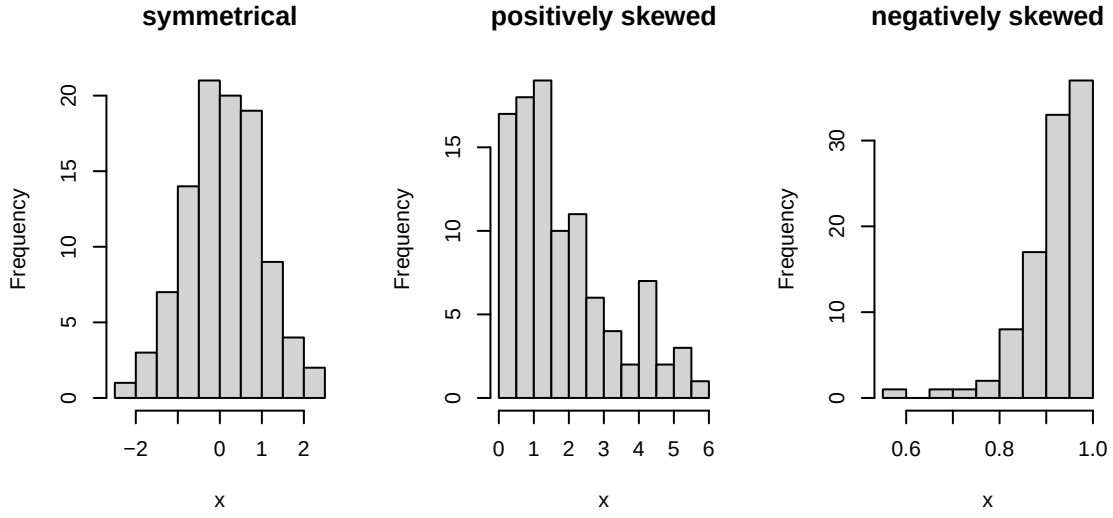
1.6 Plotting a distribution using a histogram

We've seen that 25 countries got higher scores than the UK, but perhaps some of these were only *just* higher, so that the just looking at ranks can be misleading. We can plot the distribution of scores using a **histogram**.

Definition 1.4 (Histogram)**.** A histogram is a bar chart, where the area of each bar is proportional to the number of observations lying in the interval indicated by the bar. Each interval is known as a **bin**. If the bars are all of equal width (which is recommended), then the height of the bar is normally either the number of observations in the corresponding bin, or the proportion of the total number that lie in that bin.

1.6.1 Describing the shape of a distribution: skewness

From a histogram plot, we sometimes refer to a distribution as being (approximately) symmetrical, positively skewed, or negatively skewed. Some examples are below.



Starting from the peak of a distribution, a positively skewed distribution extends further to the right than to the left, and vice-versa for a negatively skewed distribution. Note that in a positively skewed distribution, the mean is greater than the median, and vice-versa for a negatively skewed distribution.

1.7 Introducing ggplot2

We'll be using the R package `ggplot2` to produce all our plots. `ggplot2` is a popular and versatile package for producing a wide range of plots.

(Artwork by @allison_horst)

Everything you need for this module is included in these notes, but there are lots of good online resources too:

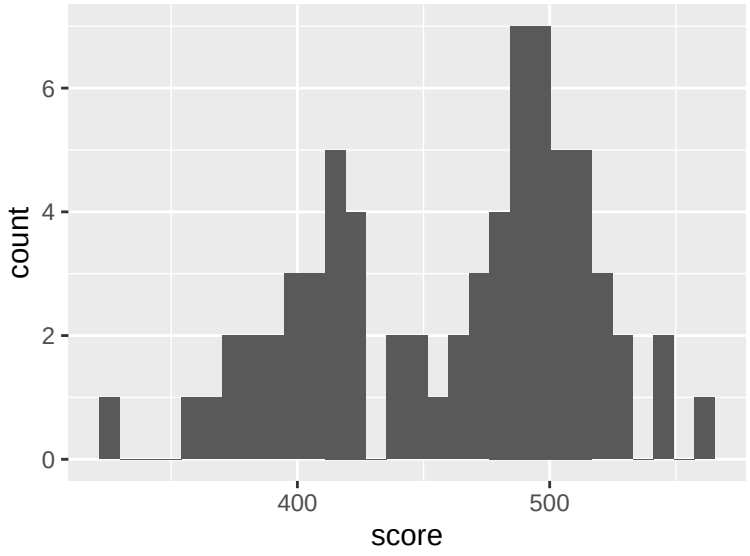
- R Studio Primers (Visualize Data)
- R for Data Science (Chapter 3)
- R Graphics Cookbook

`ggplot2` is included in the `tidyverse` package, so we don't need to install or load it separately.

1.8 Drawing a histogram in R

We can produce a basic histogram as follows

```
ggplot(data = maths, aes(x = score)) +
  geom_histogram()
```



This is really a single command, spread over two lines, using the `+` symbol to carry over the command from one line to the next year. The command works as follows.

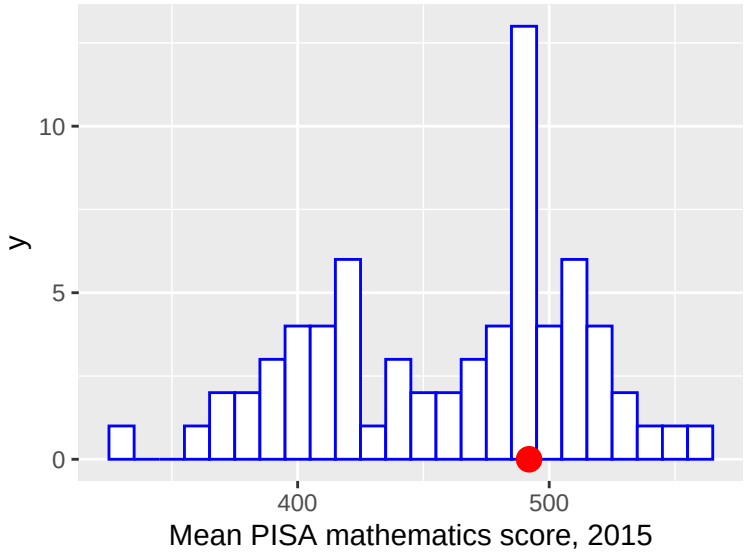
- The first line `ggplot(data = maths, aes(x = score))` sets up the axes, and tells R that we will be plotting data from the `maths` dataframe, with the dataframe column `score` represented on the x-axis.
- The second line `geom_histogram()` tells R to draw a histogram for the axes specified in the first line.

When producing any plot with `ggplot2`, any command involving a column in a dataframe needs to go **inside** an `aes()` command (e.g., `aes(x = score)`). This can take a little getting used to, but you will see more examples later on.

1.8.1 Customising a histogram plot in R

We'll redraw the plot with different colours, add a better axis label, specify a histogram bin-width of size 10, and indicate the UK's score with a red dot:

```
ggplot(data = maths, aes(x = score)) +
  geom_histogram(colour = "blue", fill = "white", binwidth = 10) +
  labs(x = "Mean PISA mathematics score, 2015") +
  annotate("point", x = 492, y = 0,  size = 4, colour = "red")
```



Note that the `+` symbol at the end of the first three lines tells R to treat all four lines as a single command.

- The second line now includes extra arguments: `fill` sets the colour of the interior of the bars, and `colour` sets the colour of the bar edges. `binwidth` sets how wide each bar is on the x-axis;
- the third line (`labs`) specifies the label on the x-axis;
- the fourth line (`annotate`) draws a red circle at the coordinates $x = 492, y = 0$, and `size = 4` increases the size of the circle (the default value for `size` is 1.)

Now we can see that quite a few countries got scores close to the UK's, so that there's not necessarily much to distinguish countries, even if they are, say, 10 places apart in the rankings.

1.9 Covariance and correlation

Let's first consider the relationship between the maths score and GDP per capita. For the i-th country, let x_i be its maths score and y_i its GDP per capita, for $i = 1, \ldots, 70$. Looking at the first few rows of the data

```
head(maths)
```

```
## # A tibble: 6 x 8
##   country         continent     score   gdp  gini homework start.age wealthiest
##   <chr>           <chr>         <dbl> <dbl> <dbl>    <dbl>     <dbl> <lgl>
## 1 Albania         Europe          413  4147  29        5.1         6 FALSE
## 2 Algeria         Africa          360  3844  27.6     NA           6 FALSE
## 3 Argentina       South America   409 12449  42.7      3.7         6 FALSE
## 4 Australia       Oceanea         494 49928  34.7      6           5 TRUE
## 5 Austria         Europe          497 44177  30.5      4.5         6 TRUE
## 6 B-S-J-G (China) Asia            531  8123  42.2     13.8         6 FALSE
```

so we have paired observations $(x_1 = 413, y_1 = 4147)$, $(x_2 = 360, y_2 = 3844)$, $(x_3 = 409, y_3 = 12449)$ and so on.

Definition 1.5 (Covariance)**.** For pairs of observations $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ we define their covariance to be

$$s_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}),$$

where $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ and $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$.

1.9.1 Calculating a covariance in R

In R, we can use the command `cov()`:

```
cov(maths$score, maths$gdp)
```

```
## [1] 710058.9
```

and so $s_{xy} = 710058.9$ to 1 d.p.

1.9.2 Pearson's correlation coefficient

Covariances aren't very informative on their own, as they will depend on the scale of measurement of the variables. **Correlation coefficients** are scale independent. There are different versions of the correlation coefficient.

Definition 1.6 (Pearson's correlation coefficient)**.** For pairs of observations $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ we define Pearson's correlation coeffcient to be

$$r_{xy} = \frac{s_{xy}}{s_x s_y},$$

with s_{xy} the covariance defined above, and

$$s_x = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2},$$

$$s_y = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

Pearson's correlation coefficient measures the strength of the *linear* association between the two variables, and is bounded between -1 and 1. A positive correlation implies that as one quantity increases, the other is expected to increase, and a negative correlation implies that as one quantity increases, the other is expected to *decrease*.

1.9.3 Calculating Pearson's correlation coefficient in R

To calculate Pearson's correlation coefficient between the variables `score` and `gdp` in the data frame `maths`, we use the command

```
cor(maths$score, maths$gdp)
```

```
## [1] 0.5947452
```

and so $r_{xy} = 0.59$ to 2 d.p. We will discuss interpreting correlation coefficients shortly, but for now, we will just comment that this value is fairly large, and certainly suggests a relationship between the two variables.

1.9.4 Spearman's correlation coefficient

An alternative to Pearson's correlation coefficient is **Spearman's** correlation coefficient.

Definition 1.7 (Spearman's correlation coefficient)**.** For pairs of observations $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ we define Spearman's correlation coefficient to be Pearson's correlation coefficient calculated on the the *ranks* of observations.

1.9.4.1 Example

For illustration, suppose we have the following data

i	1	2	3	4	5	6
x_i	68	2	40	20	85	97
y_i	73	26	37	1	63	68

We first calculate the ranks of the observations (if x_i gets a rank of 1, it means x_i was the smallest out of $x_1, \ldots, x_n$):

i	1	2	3	4	5	6
rank(x_i)	4	1	3	2	5	6
rank(y_i)	6	2	3	1	4	5

We then calculate Pearson's correlation coefficient on the ranks, as if we have six pairs of observations (4, 6), (1, 2), ... (6, 5).

1.9.5 Calculating Spearman's correlation coefficient in R

In R, we just include an extra argument in the `cor()` command:

```
x <- c(68, 2, 40, 20, 85, 97)
y <- c(73, 26, 37, 1, 63, 68)
cor(x, y, method = 'spearman')
```

```
## [1] 0.7714286
```

(If we don't specify a `method`, the default is to use Pearson's). To illustrate that this is just Pearson's correlation coefficient calculate on the ranks, we can obtain the rankings in R with the command `rank()`, and then compare the above with

```
rank(x)
```

```
## [1] 4 1 3 2 5 6
```

```
rank(y)
```

```
## [1] 6 2 3 1 4 5
```

```
cor(rank(x), rank(y))
```

```
## [1] 0.7714286
```

1.9.6 Interpreting correlation coefficients

To help interpret and visualise correlation coefficients, four examples are plotted in Figure 1.1.

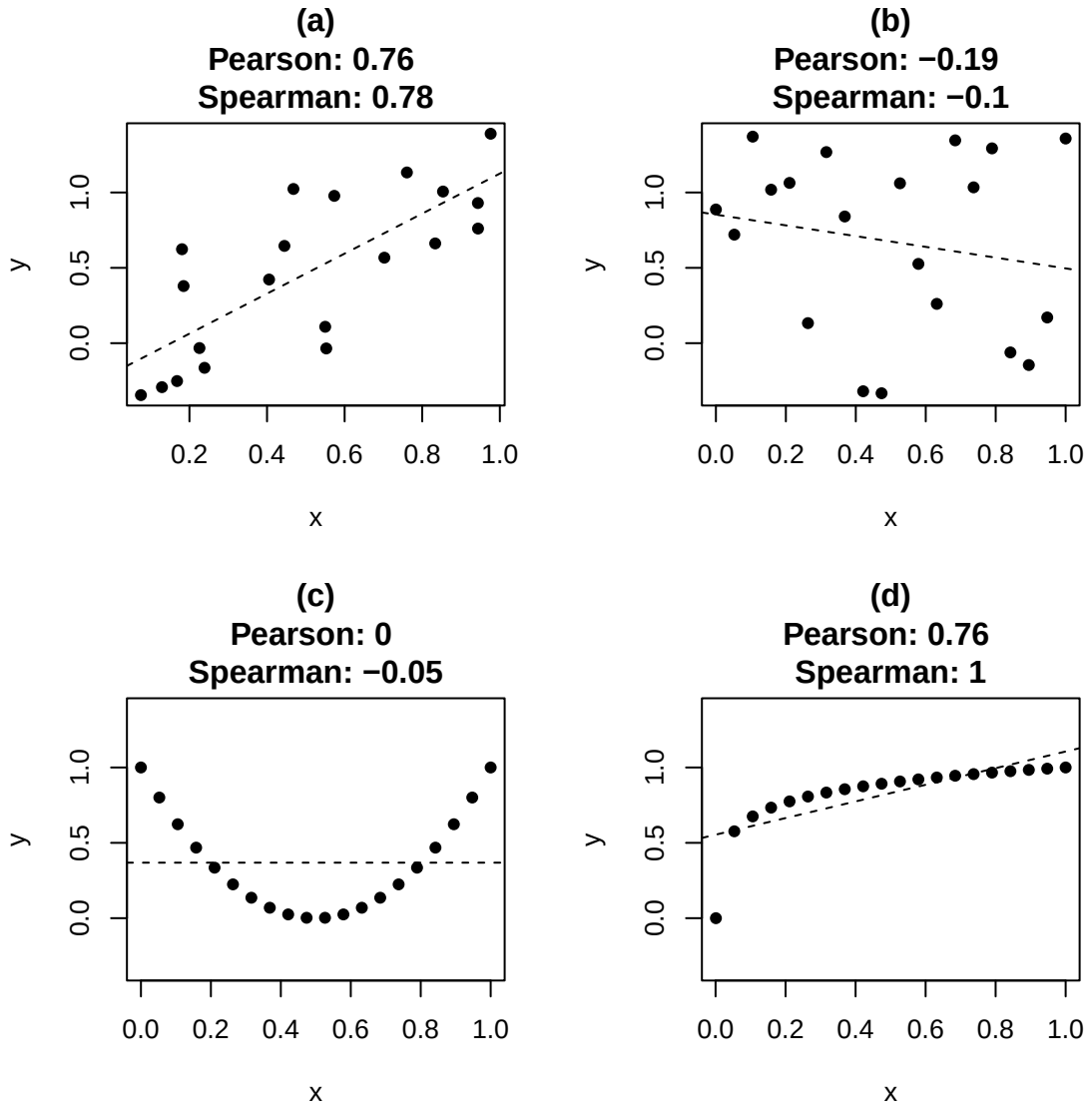


Figure 1.1: Correlation coefficients and linear trends. In (a), the correlation is quite large, even though there is fair amount of 'random variation' in the data. In (b), the two variables were generated independently, but still resulted in a Pearson correlation of -0.19: 'moderate' correlation values can be observed purely by luck. In (c), x and y are clearly related, but the Pearson correlation is 0: there is no *linear* trend. In (d), the relationship between x and y is monotone, but nonlinear, and the Spearman correlation is greater than Pearson's.

1.9.7 Correlations for the `maths` data set

We calculate the Pearson correlations between the variables of interest:

```
maths %>%
  na.omit() %>%
  select(score, gdp, homework, start.age, gini) %>%
  cor(method = "pearson") %>%
  round(2)
```

- the second line excludes any rows with missing values (the `cor()` command won't work otherwise);
- the third line selects the columns for which we wish to calculate correlations;
- the fourth line will produce a matrix of Pearson correlations, in the form above;
- the fifth line rounds all the numbers to two decimal places.

The code above produces the following output.

```
##            score   gdp homework start.age  gini
## score       1.00  0.58     0.26      0.06 -0.56
## gdp         0.58  1.00    -0.14     -0.19 -0.40
## homework    0.26 -0.14     1.00      0.13  0.07
## start.age   0.06 -0.19     0.13      1.00 -0.11
## gini       -0.56 -0.40     0.07     -0.11  1.00
```

(Note that the correlation of any variable with itself is always 1). By looking at the largest correlations (in absolute value), tentatively, we may conclude the following:

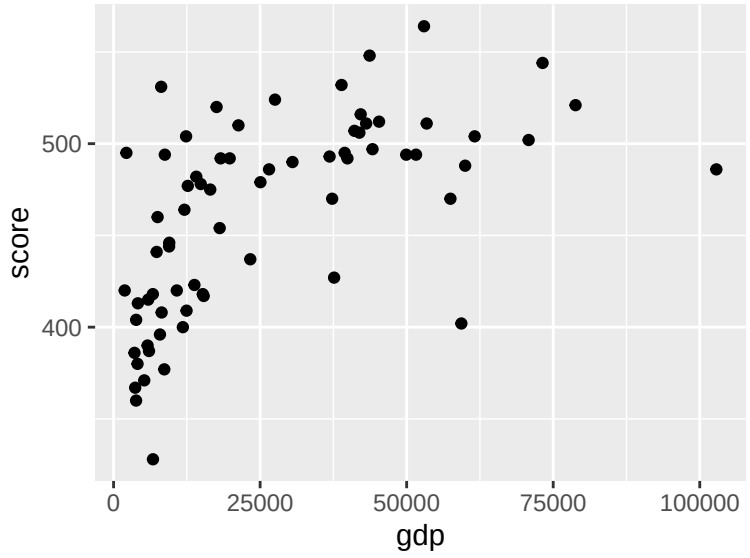
- countries with larger GDP per capita tend to have higher maths scores (correlation of 0.58);
- countries with more inequality tend to have lower maths scores (correlation of -0.56);
- the association between hours of homework per week and maths score looks weak (correlation of 0.26);
- there doesn't appear to be much of a linear association between school starting age and maths score (correlation of 0.06).

(The correlations are all fairly similar if we use Spearman's correlation instead). We will check these visually shortly. There is one complication, however: countries with higher GDP per capita tend to have less inequality (correlation of -0.4). Could this be the reason countries with more inequality tend to have lower maths scores?

1.10 Drawing a scatter plot in R

Let's first plot maths score against GDP per capita:

```
ggplot(data = maths, aes(x = gdp, y = score)) +
  geom_point()
```



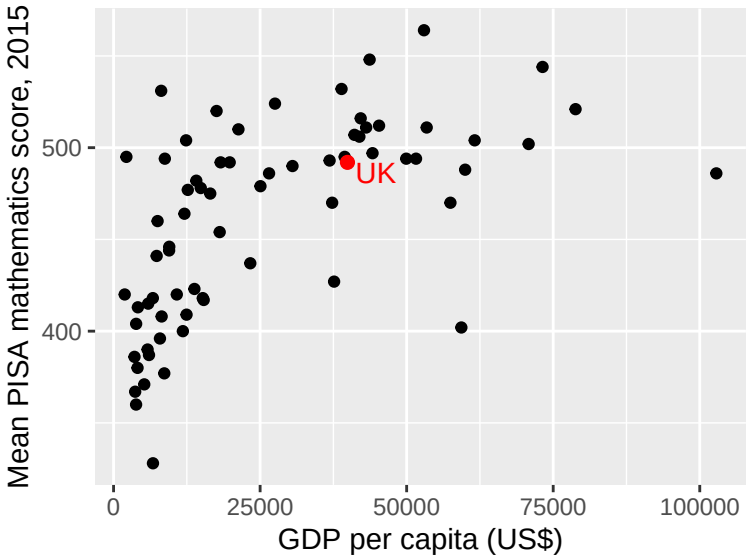
Again, although there are two lines, with the `+` symbol, this is really just one command.

- The first line `ggplot(data = maths, aes(x = gdp, y = score))` sets up the axes, and tells R that we will be plotting data from the `maths` dataframe, with the dataframe column `gdp` represented on the x-axis, and the column `score` represented on the y-axis.
- Both `gdp` and `score` are columns in a data frame, so they are used inside an `aes()` command here.
- The second line `geom_point()` tells R to draw a scatter plot for the axes specified in the first line.

1.10.1 Customising a scatter plot in R

We will tidy up the plot by specifying proper axes labels, and labeling the UK. Here, the simplest way to do it is to manually add another point.

```
ggplot(data = maths, aes(x = gdp, y = score)) +
  geom_point() +
  labs(x = "GDP per capita (US$)", y = "Mean PISA mathematics score, 2015") +
  annotate("point", x = 39899, y = 492, colour = "red", size = 2) +
  annotate("text", label = "UK", x = 39899, y = 492, colour = "red",
           hjust = -0.2, vjust = 1)
```



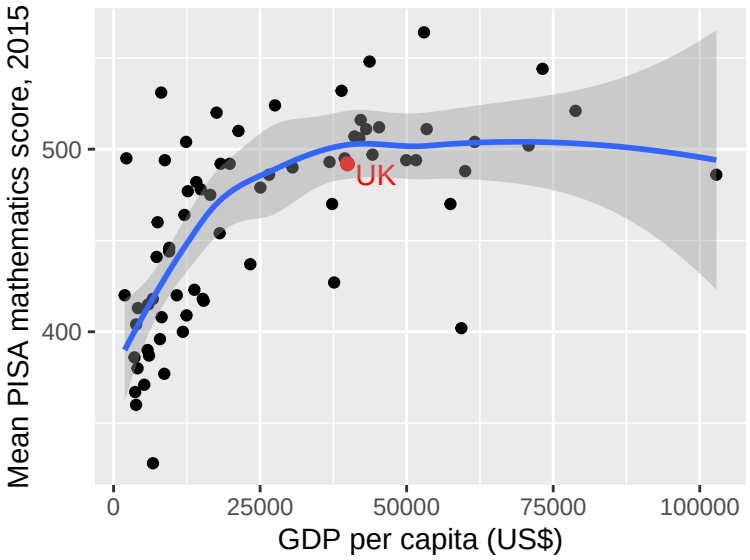
Note that you can break up long lines of code after a comma `,` which makes your code easier to read.

- The third line (`labs`) specifies labels for both the x-axis and y-axis;
- the fourth line adds a red circle (`"point"`) at the coordinates $x = 39899$, $y = 492$ (corresponding to the UK), with the `size` set to 2 to make it a little larger.
- the fifth line adds some red text (UK) at the coordinates $x = 39899$, $y = 492$, with the arguments `hjust` and `vjust` shifting the text slightly horizontally and vertically, so that it appears next to, rather than on top of the red dot. It can take a little trial and error to find the values for `hjust` and `vjust` that you are happy with.

1.10.2 Adding a nonlinear trend to a scatter plot in R

We can see clearly that maths scores tend to increase as GDP per capita increases, but the relationship doesn't look linear. If we want to emphasise such a relationship, we can add the trend to the plot, using the extra line `geom_smooth()` in the plot command:

```
ggplot(data = maths, aes(x = gdp, y = score)) +
  geom_point() +
  labs(x = "GDP per capita (US$)",
       y = "Mean PISA mathematics score, 2015") +
  annotate("point", x = 39899,
           y = 492, colour = "red",
           size = 2) +
  annotate("text", label = "UK",
           x = 39899, y = 492,
           colour = "red",
           hjust = -0.2, vjust = 1) +
  geom_smooth()
```

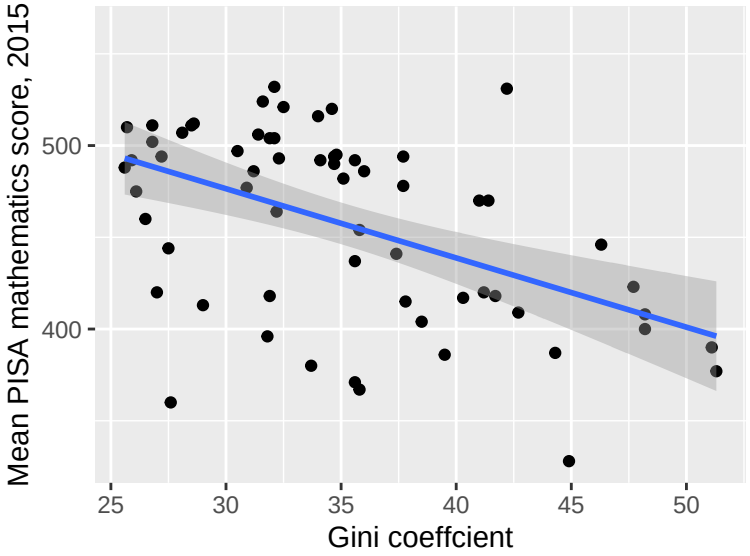


The blue line shows the estimated trend. The grey shaded area indicates uncertainty about this trend: it's wider on the right hand side, because we have less data there. (You can learn more about how this is all done in MPS235).

1.10.3 Adding a linear trend to a scatter plot in R

To include a linear trend, we add the argument `method = "lm"` to `geom_smooth()`:

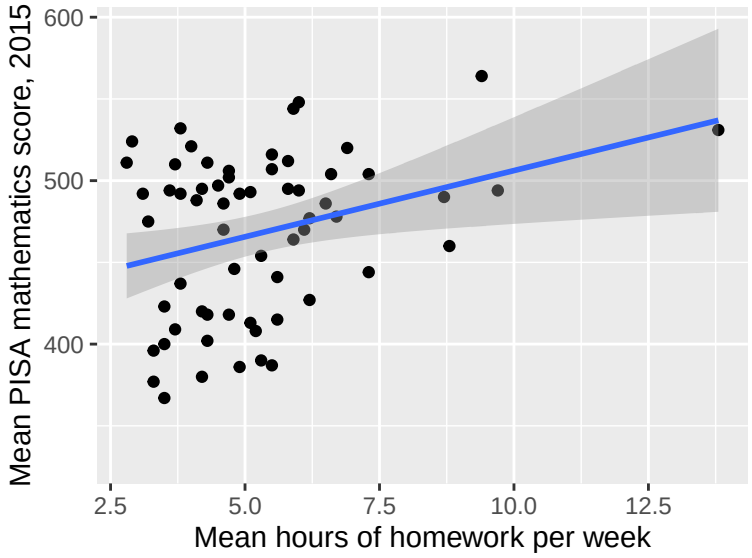
```
ggplot(data = maths,
       aes(x = gini, y = score)) +
  geom_point() +
  geom_smooth(method = "lm") +
  labs(x = "Gini coeffcient",
       y = "Mean PISA mathematics score, 2015")
```



The grey shaded region represents uncertainty in the trend. We'll do the same for the homework

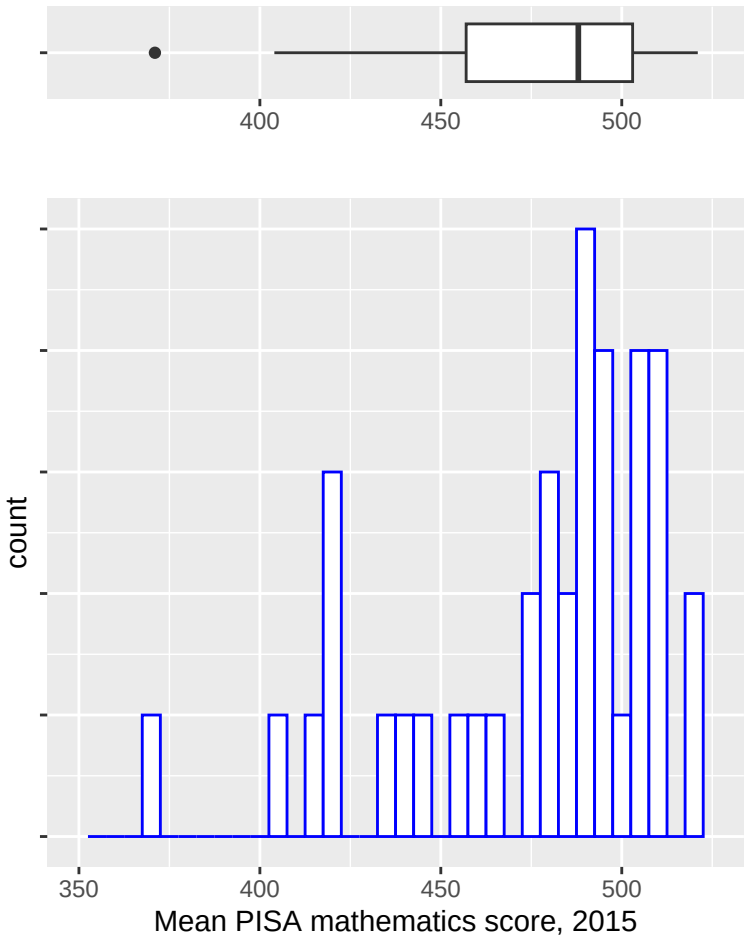
variable.

```
ggplot(data = maths,
       aes(x = homework, y = score)) +
  geom_point() +
  geom_smooth(method = "lm") +
  labs(x = "Mean hours of homework per week")+
  labs(y = "Mean PISA mathematics score, 2015")
```



1.11 Box plots

If we want to compare groups of observations, we can use a box plot. Box plots are useful for comparing several groups simultaneously: we can fit a lot of information into a single plot. To explain how to interpret a box plot, we'll first consider the European countries only, and show a box plot directly above a histogram of the same data.



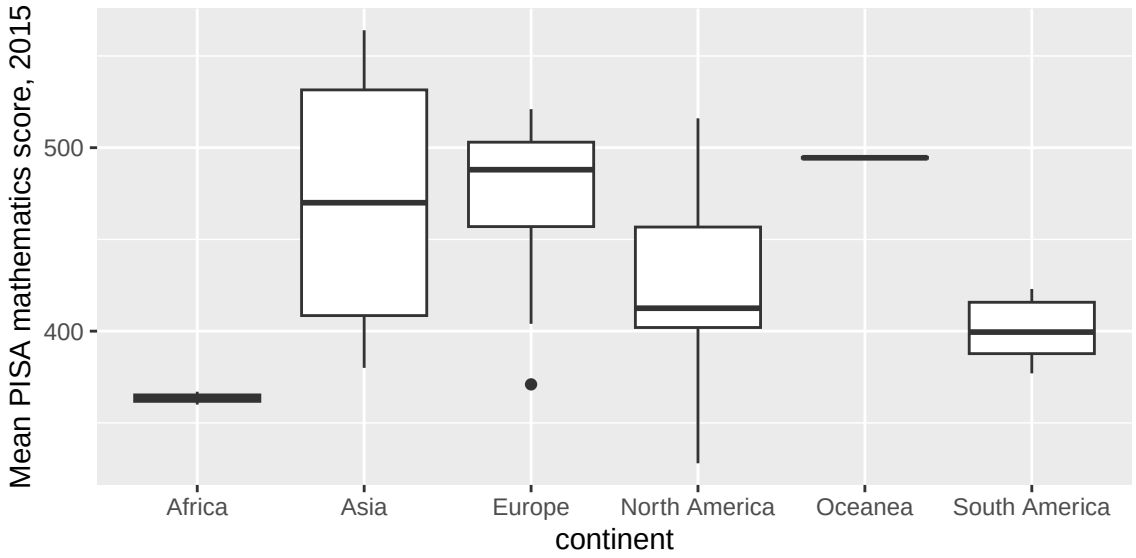
In the box plot above:

- the thick vertical black line shows the median;
- the ends of the box show the 25th and 75th percentiles (so that the box shows the interquartile range);
- the two horizontal lines (known as "whiskers") extend to the most extreme observed values that are no more than 1.5 × the inter-quartile range from the edge of the box;
- individual observations not covered by the whiskers are shown as points on the plot, referred to as **outliers**.

Lining up the box plot with the histogram, we can see that it gives an indication of the shape of the distribution.

We'll now produce a box plot (drawn vertically instead of horizontally) to compare maths scores between the continents.

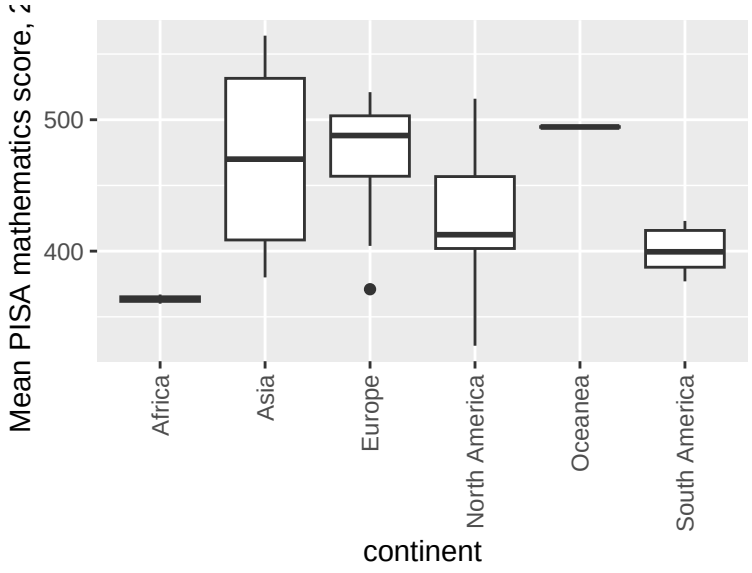
```
ggplot(data = maths, aes(x = continent, y = score)) +
  geom_boxplot() +
  labs(y = "Mean PISA mathematics score, 2015")
```



- The first line `ggplot(data = maths, aes(x = continent, y = score))` sets up the axes, and tells R that we will be plotting data from the `maths` dataframe, with the dataframe column `continent` represented on the x-axis, and the column `score` represented on the y-axis.
- The second line `geom_boxplot()` tells R to draw a box plot for the axes specified in the first line.

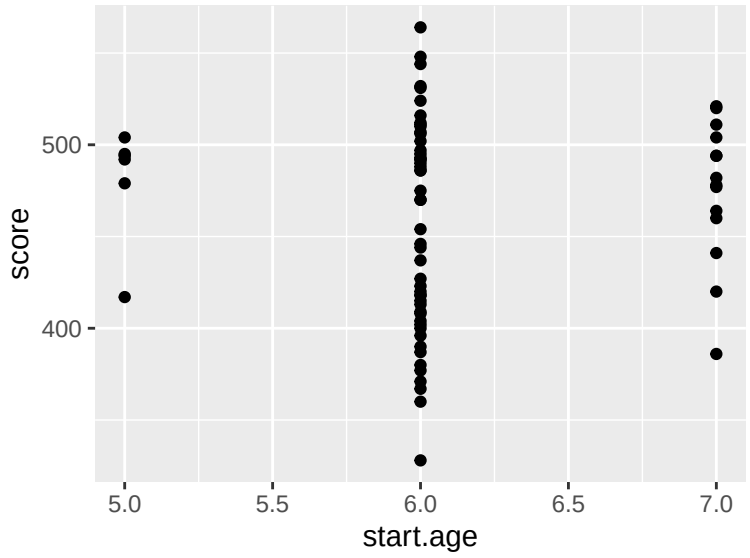
If we have long names on the x-axis, we may need to rotate them. We can do this as with an extra line at the end (which will rotate and shift the text slightly):

```
ggplot(data = maths, aes(x = continent, y = score)) +
  geom_boxplot() +
  labs(y = "Mean PISA mathematics score, 2015") +
  theme(axis.text.x=element_text(angle = 90,
                                 hjust = 1,
                                 vjust = 0.5))
```



Let's now investigate the relationship between maths score and school starting age:

```
ggplot(data = maths, aes(x = start.age, y = score)) +
  geom_point()
```



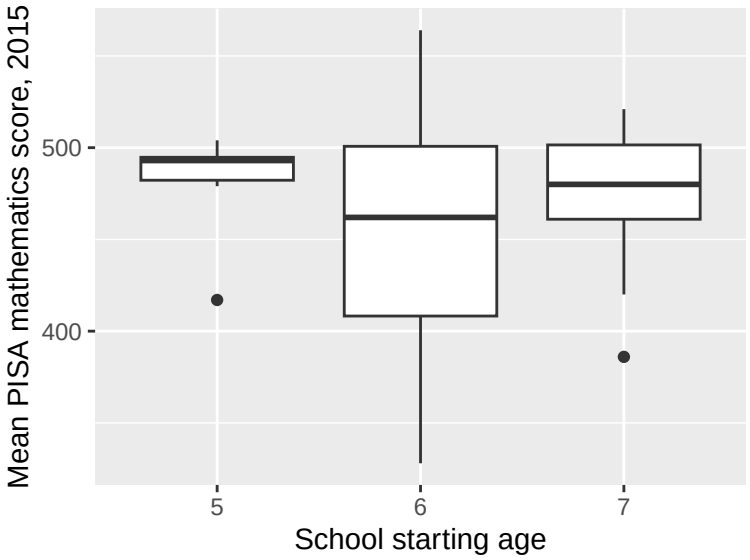
The problem here is that there are only three distinct starting ages: 5, 6 and 7, so the scatter plot isn't very clear. A box plot might work better: we think of the three starting ages as three groups. In R, we have to convert `start.age` to a `factor` variable. If we try the following command

```
factor(maths$start.age)
```

```
##  [1] 6 6 6 5 6 6 6 6 7 6 6 6 6 7 6 6 6 6 7 7 6 6 6 6 6 7 6 7 5 6 6 6 6 7 6 7 6 7
## [39] 6 6 6 6 5 6 7 6 6 5 6 6 7 6 6 6 7 6 6 6 6 7 7 6 5 6 6 6 5 6 6 6
## Levels: 5 6 7
```

this doesn't appear to have done anything, but the text `Levels: 5 6 7` at the bottom tells us that the output is a factor variable, taking one of three levels. We can try the box plot, where on the x-axis, we specify that we want to use the factor variable `factor(start.age)`:

```
ggplot(data = maths, aes(x = factor(start.age), y = score)) +
  geom_boxplot() +
  labs(x = "School starting age",
       y =  "Mean PISA mathematics score, 2015")
```



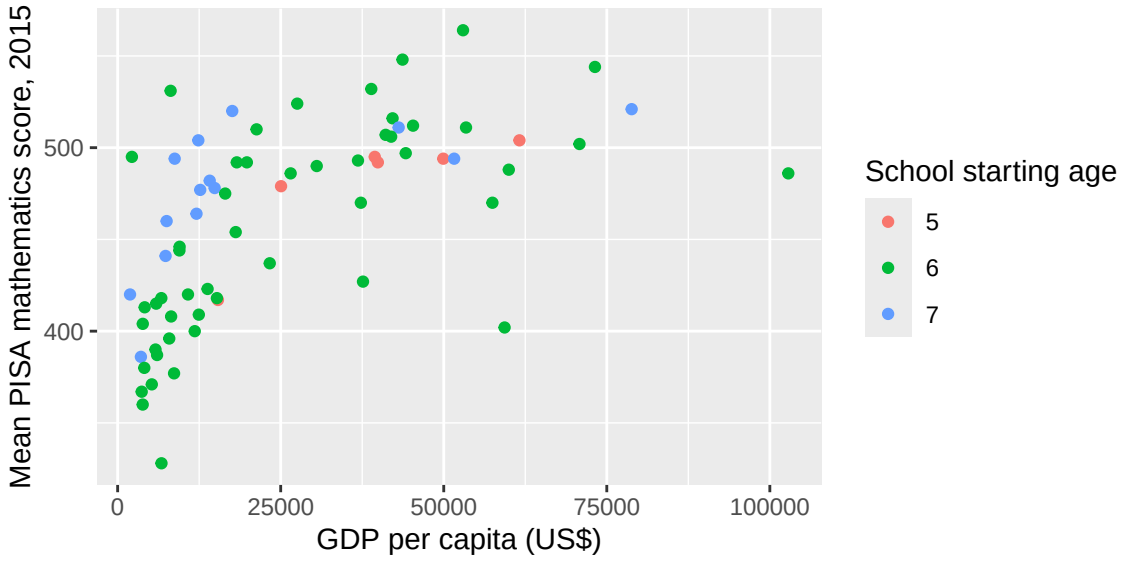
The median score is highest in those countries where children start school at age 5, but we can see there is a lot of variation within each group of countries.

1.12 Scatter plots to represent three variables

Sometimes, we may wish to visualise the relationship between several variables simultaneously. One way to do this is with a scatter plot, using the colour of each point to represent a third variable.

We'll plot score against GDP as before, but now indicate the school starting age with different colours (after first converting `start.age` to a factor variable):

```
ggplot(data = maths,
       aes(x = gdp, y = score)) +
  geom_point(aes(colour = factor(start.age))) +
  labs(x = "GDP per capita (US$)",
       y = "Mean PISA mathematics score, 2015") +
  labs(colour = "School starting age")
```



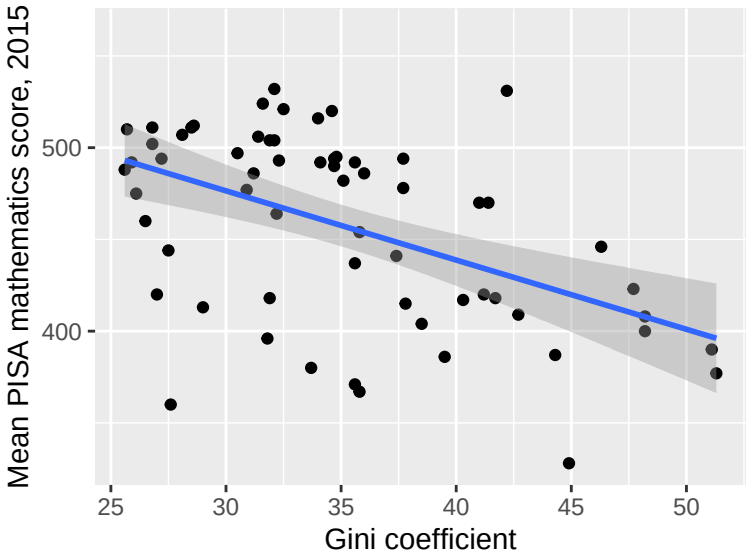
- we want `start.age` to be treated as a factor, so we convert it to a factor with `factor(start.age)`
- in the `geom_point()` command, we want the `colour` of the points to be determined by `factor(start.age)`
- `factor(start.age)` refers to a column in a data frame, so it must be used inside an `aes()` command.
- the command `labs(colour = "School starting age")` sets the title for plot legend, where the legend explains what `colour` represents in the scatter plot.

Some conclusions:

- in most of the poorer countries, children either start school at age 6 or 7: this has brought down the median compared with starting age 5 group;
- within the poorer countries, scores do appear to be higher where children have started at age 7;
- within the wealthier countries, there is no obvious effect of school starting age.

Now we'll consider the effect of income inequality. We first do a simple scatter plot with a linear trend

```
ggplot(data = maths,  aes(x = gini, y = score)) +
  geom_point() +
  geom_smooth(method = "lm") +
  labs(x = "Gini coefficient",
       y = "Mean PISA mathematics score, 2015")
```

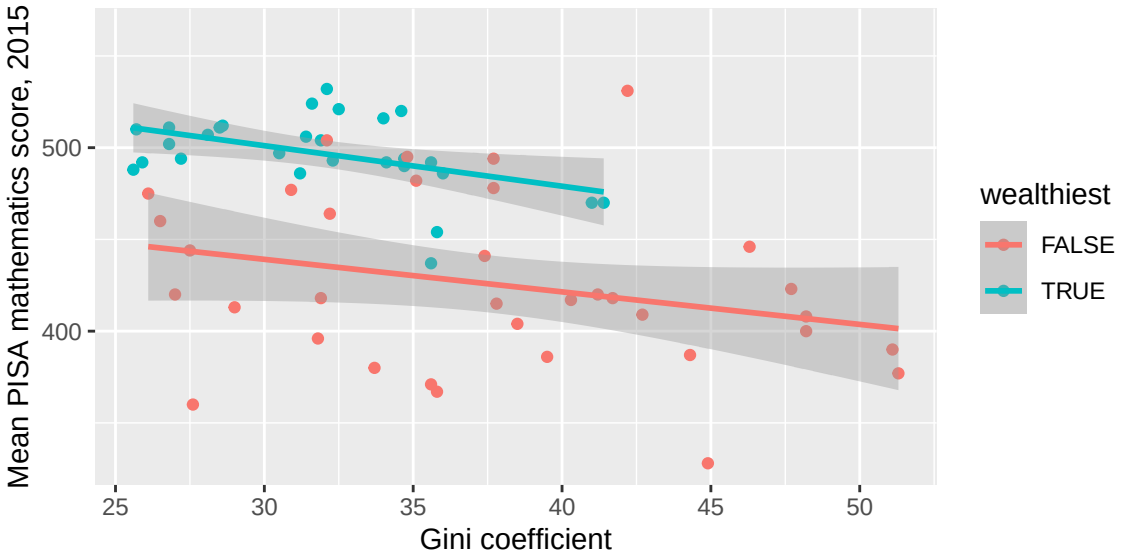


It does appear that score decreases with increasing income inequality (as we saw when we calculated the correlations). We also noted before that income inequality is negatively correlated with GDP. We would like to take account for this. Recall that the new variable `wealthiest` indicates whether each country is in the top half of the wealthiest countries, as measured by GDP.

```
## # A tibble: 6 x 5
##   country         score   gdp  gini wealthiest
##   <chr>           <dbl> <dbl> <dbl> <lgl>
## 1 Albania           413  4147  29   FALSE
## 2 Algeria           360  3844  27.6 FALSE
## 3 Argentina         409 12449  42.7 FALSE
## 4 Australia         494 49928  34.7 TRUE
## 5 Austria           497 44177  30.5 TRUE
## 6 B-S-J-G (China)   531  8123  42.2 FALSE
```

One possibility is to draw another scatter plot, but this time using the variable `wealthiest` to determine both the colour and to display separate fitted trends: we look to see if the relationship between income equality and maths score is still there within each group:

```
ggplot(data = maths,  aes(x = gini, y = score)) +
  geom_point(aes(colour = wealthiest))+
  geom_smooth(aes(group = wealthiest,
                  colour = wealthiest),
              method = "lm") +
  labs(x = "Gini coefficient",
       y = "Mean PISA mathematics score, 2015")
```



- in the `geom_point` command, we want the `colour` of the points to be determined by the value of `wealthiest`;
- `wealthiest` is a column in a dataframe, so the argument `colour = wealthiest` has to go inside an `aes()` command;
- in the `geom_smooth` command, we want separate trends displayed for the two groups determined by the variable `wealthiest`, so we specify `group = wealthiest` inside an `aes()` command;
- in the `geom_smooth` command, we also want the lines to have different colours, so the colours are specified via `colour = wealthiest` inside an `aes()` command.
- the argument `method = lm` specifies that a linear trend should be displayed.

The gradients aren't quite as steep, but the relationship is still there. A more formal statistical modelling approach can also be used to investigate this: you can learn about this in MPS235.

Chapter 2

Machine Learning

The term "Machine Learning" was used in a paper by Arthur Samuel in 1959[1], on how a computer could learn to play a better game of checkers. In general machine learning involves getting a computer to 'learn from experience', so that it can perform a particular task in a new situation without being programmed what to do. Machine learning underpins technology described as "artificial intelligence".

Some machine learning problems involve getting computers to recognise speech, read handwriting, and identify objects in images. More recently, machine learning methods have been used to analyse very large and complex data sets, where the scale of the problem is too large for humans to manage (e.g. recommending products to millions of individual customers, based on vast databases of customer purchases).

This chapter will give a short introduction, where we will look at a single problem and method, and we will consider the role of data analysis within machine learning.

2.1 Can we teach a computer to identify handwritten digits?

Below are some scanned images of handwritten digits. The images are from the (well-known) "MNIST" data set, hosted on Yann Lecun's website.

We can see what the digits are without too much difficulty, but can we get a computer to recognise the digits?

[1]Samuel, Arthur (1959). Some Studies in Machine Learning Using the Game of Checkers. IBM Journal of Research and Development. 3 (3): 210-229.

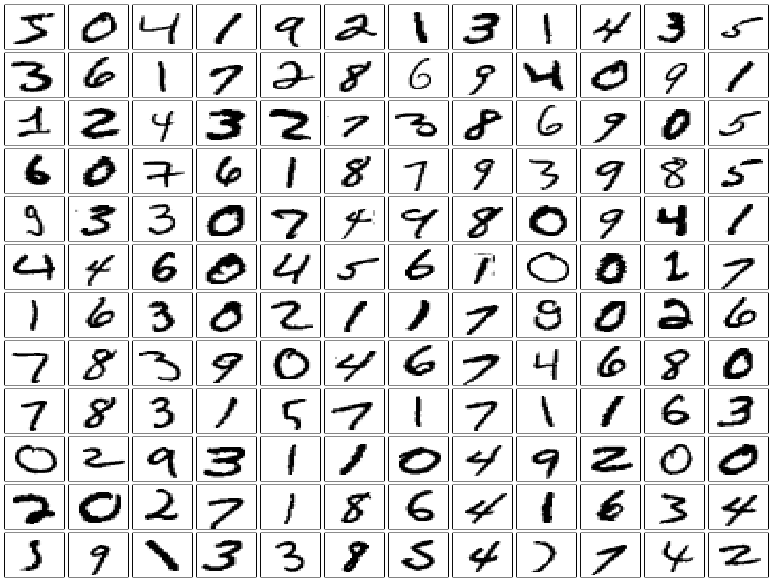


Figure 2.1: Scanned images of hand-written digits (not all written by the same person). We can easily recognise what the digits are: could a computer do so?

The key idea is to make this a *data analysis* problem. The steps are as follows.

1. Convert the objects we want the computer to recognise (the scanned images) into numerical data: find a way to represent each image with a set of numbers. The computer will need to be able to do this conversion automatically, *without* knowing what each digit in a scanned image actually is.
2. Construct a **training data set**: a (large) data set of example handwritten digits, all converted into numerical data. In addition, we *also* tell the computer what each handwritten digit is: "The first image in the data set is the number 5, the second image in the number 0," and so on.
3. Construct a statistical model or algorithm using the training data set, that given an image in its converted numerical form, can estimate what the handwritten digit is.

In this module, steps 1 and 2 will always be done for you; you will only need one method for doing step 3.

2.1.1 Step 1: converting an image into data

(The images first need to be scaled to the same size, with the handwriting approximately in the centre of the image. This step has been done for us.)

Here is one example image:

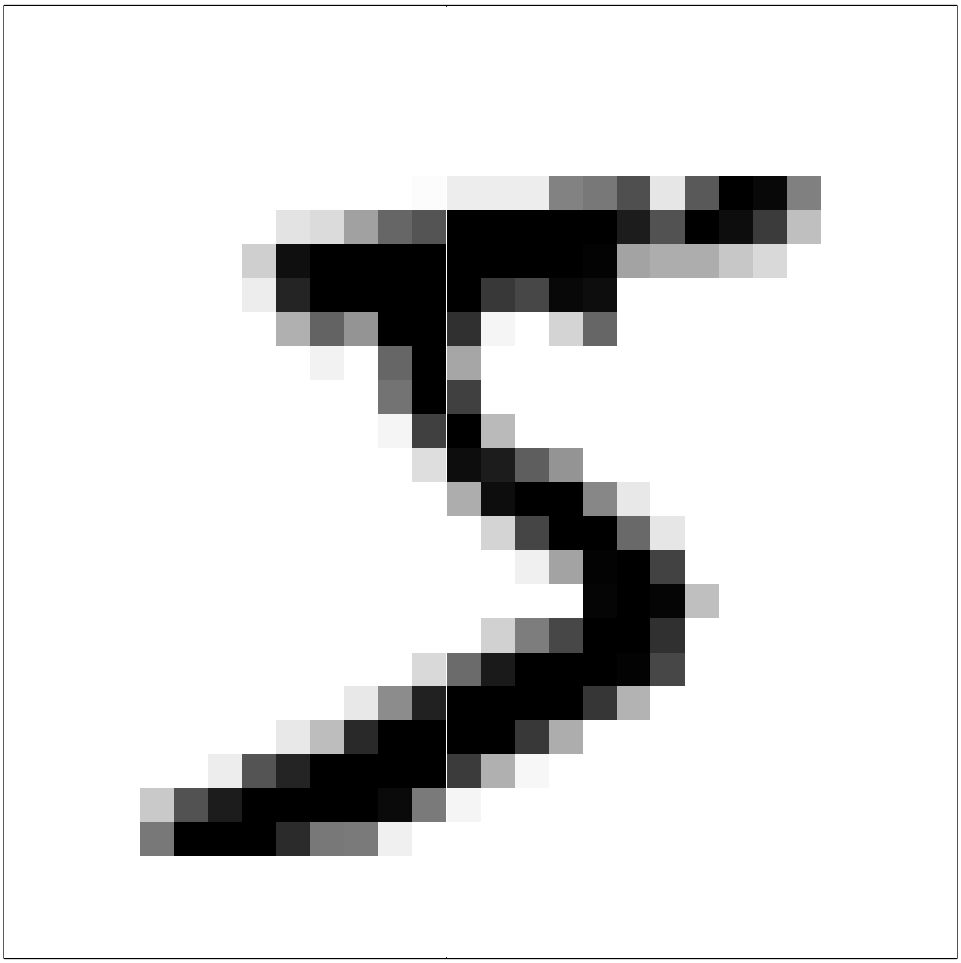


Figure 2.2: A single scanned image of a handwritten digit. (The top left corner image from Figure 2.1.) Zooming in, we can see that the image is made up of shaded blocks.

The image is made up of pixels (shaded dots), arranged on a 28x28 grid (these are low resolution images!) The shading of each pixel can be represented numerically, on a scale of 0 (white) to 255 (black).

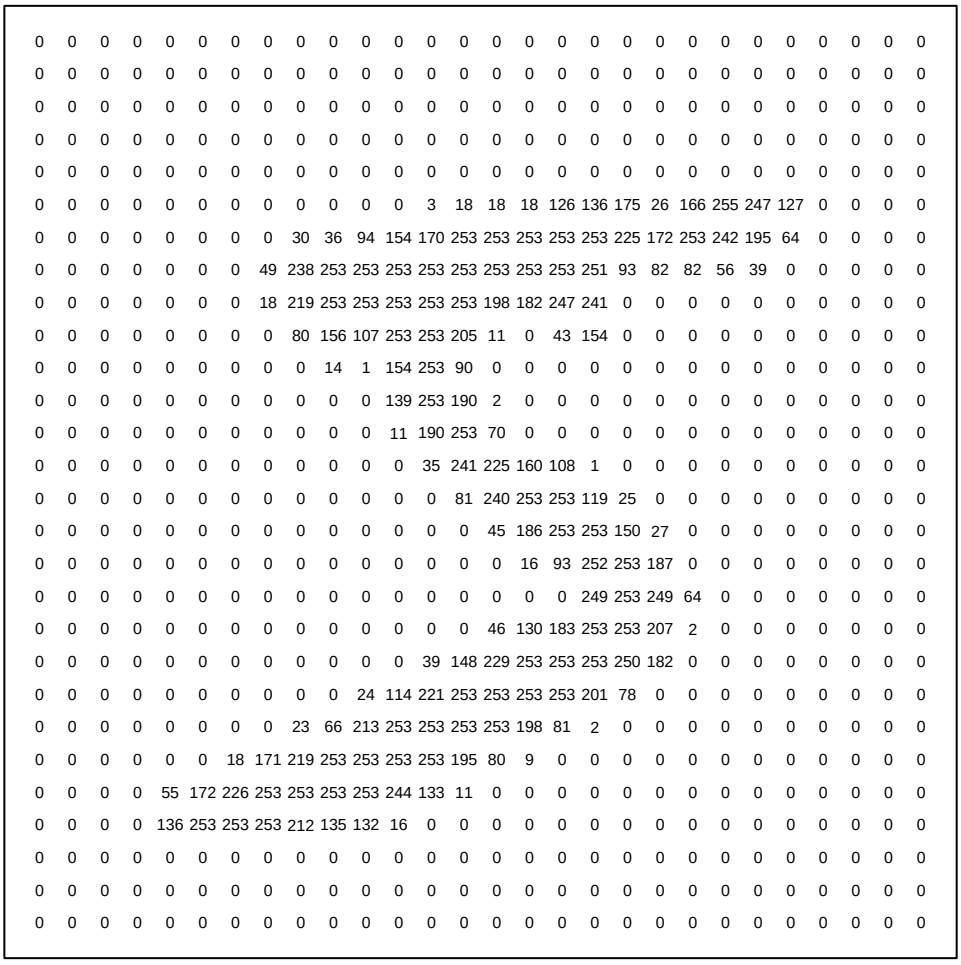


Figure 2.3: This is how the computer 'sees' the image, as a grid of pixel shades (here with higher numbers representing darker shades). We can use these numbers to represent the image in numerical form.

We can now represent the image by a vector $\mathbf{x}$ with $28^2 = 784$ elements:

$$\mathbf{x} = (x_1, x_2, \dots, x_{784}),$$

taking one row at a time from the above image. Starting with the top row, the vector would look like this:

$$\mathbf{x} = (0, 0, 0, \dots, 0, 0, 0, 3, 18, 18, 18, 126, 136, \dots, 135, 132, 16, 0, 0, \dots, 0, 0, 0)$$

2.1.2 Step 2: assembling the training data set

We now assemble a training data set of 60,000 images. The i-th image in the data set is represented by the vector $\mathbf{x}_i$, where

$$\mathbf{x}_i = (x_{i,1}, x_{i,2}, \dots, x_{i,784})$$

We define y_i to be the **class label** for the i-th image. In this case, the class label will be a number from 0-9, that says what the hand written digit is. For the training data set, the class labels are simply given to us: we don't need to work out what they should be. We write the training data set

in the form

$$(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \ldots, (\mathbf{x}_{60,000}, y_{60,000}).$$

If $(\mathbf{x}_1, y_1)$ corresponds to the first image in the training data set, then $\mathbf{x}_1$ is the vector with elements given by the numbers in Figure 2.3, and $y_1 = 5$: the image is a hand-written number 5.

2.1.2.1 Setting up the data in R.

If you want to get the data for yourself, I suggest you download the data in csv format: csv files `mnist_train.csv` and `mnist_test.csv` can be downloaded from this site maintained by Joseph Redmon. Once you have downloaded them, assuming the files are in your working directory, use the commands

```
library(tidyverse)
training_set <- read_csv("mnist_train.csv", col_names = FALSE)
test_set <- read_csv("mnist_test.csv", col_names = FALSE)
```

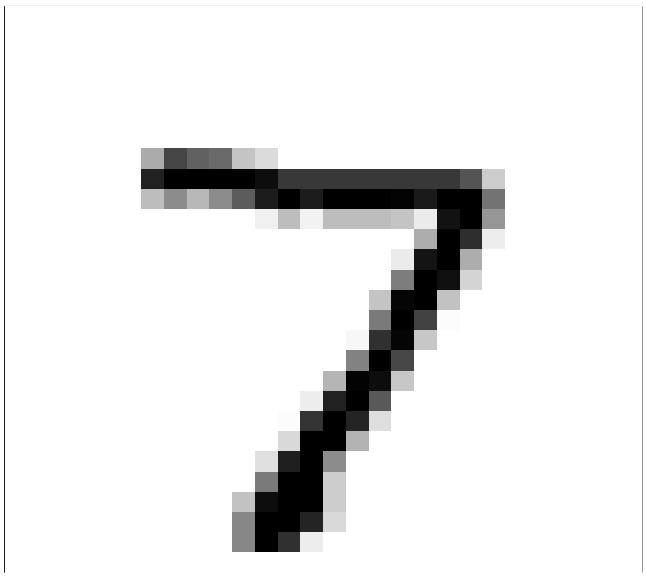
For the training data, the class labels $y_1, y_2, \ldots, y_{60000}$ are stored in the first column of `training_set`. The image vector x_i is stored in row i, columns 2 to 725. As well as the training data, we have another data set know as the "test data", which are stored are stored similarly in `test_set`.

2.1.3 Step 3: an algorithm for estimating the digit in a new image

We have a separate **test data set** made up of 10,000 images. We will represent these by the vectors

$$\tilde{\mathbf{x}}_1, \tilde{\mathbf{x}}_2, \ldots, \tilde{\mathbf{x}}_{10,000}$$

The first image in our test data set and its numerical representation $\tilde{\mathbf{x}}_1$ are shown below.



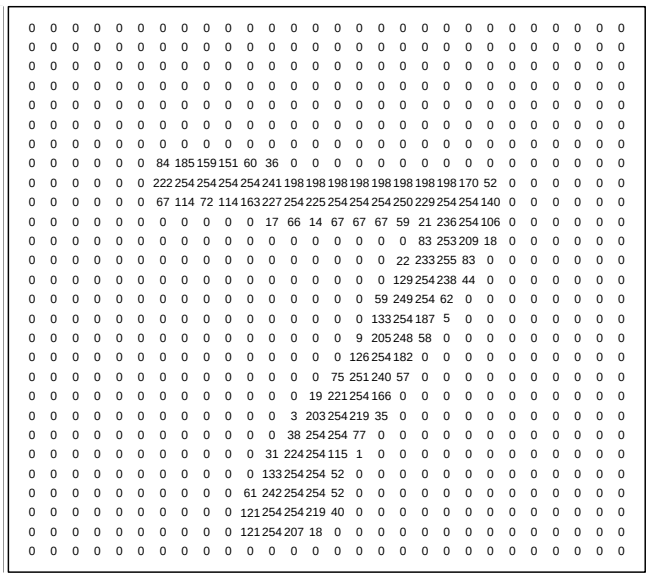


Figure 2.4: The first image in the test data set, represented by a vector $\tilde{\mathbf{x}}_1$, with elements made up of the numbers on the right hand side. How can a computer use the training data to tell that this is a number '7'?

How to get the computer to recognise that this is a number 7? There are lots of algorithms we could try (and, in general, many statistical models can be used for machine learning). Here, we will use a simple one, known as "K nearest neighbours".

2.1.4 The K nearest neighbour algorithm (KNN)

In the K nearest neighbours method, the computer will decide what number an image is by looking for similar images in the training data. It can then use the known class labels in the training images to estimate what the new test image is. Writing out the vector

$$\tilde{\mathbf{x}}_1 = (\tilde{x}_{1,1}, \tilde{x}_{1,2}, \ldots, \tilde{x}_{1,784})$$

we will measure similarity using the (square of) the Euclidean distance between $\tilde{\mathbf{x}}_1$ and each training image $\mathbf{x}_i$. We define

$$d(\tilde{\mathbf{x}}_1, \mathbf{x}_i) := \sum_{j=1}^{784} (\tilde{x}_{1,j} - x_{i,j})^2,$$

so the smaller the distance, the more similar the images are. The computer will decide what digit $\tilde{\mathbf{x}}_1$ is as follows:

1. Compute $d(\tilde{\mathbf{x}}_1, \mathbf{x}_i)$ for $i = 1, 2, \ldots, 60000$
2. Find the nearest neighbour: look for smallest distance out of

$$d(\tilde{\mathbf{x}}_1, \mathbf{x}_1), \quad d(\tilde{\mathbf{x}}_1, \mathbf{x}_2), \ \ldots \ , d(\tilde{\mathbf{x}}_1, \mathbf{x}_{60000})$$

3. If the nearest neighbour was image j (the smallest distance was $d(\tilde{\mathbf{x}}_1, \mathbf{x}_j)$), then estimate the digit to be y_j: the known class label for image j.

As an example, we compute

$$d(\tilde{\mathbf{x}}_1, \mathbf{x}_1) = \sum_{j=1}^{784} (\tilde{x}_{1,j} - x_{1,j})^2 \tag{2.1}$$

$$= 5,739,837 \tag{2.2}$$

We view this below.

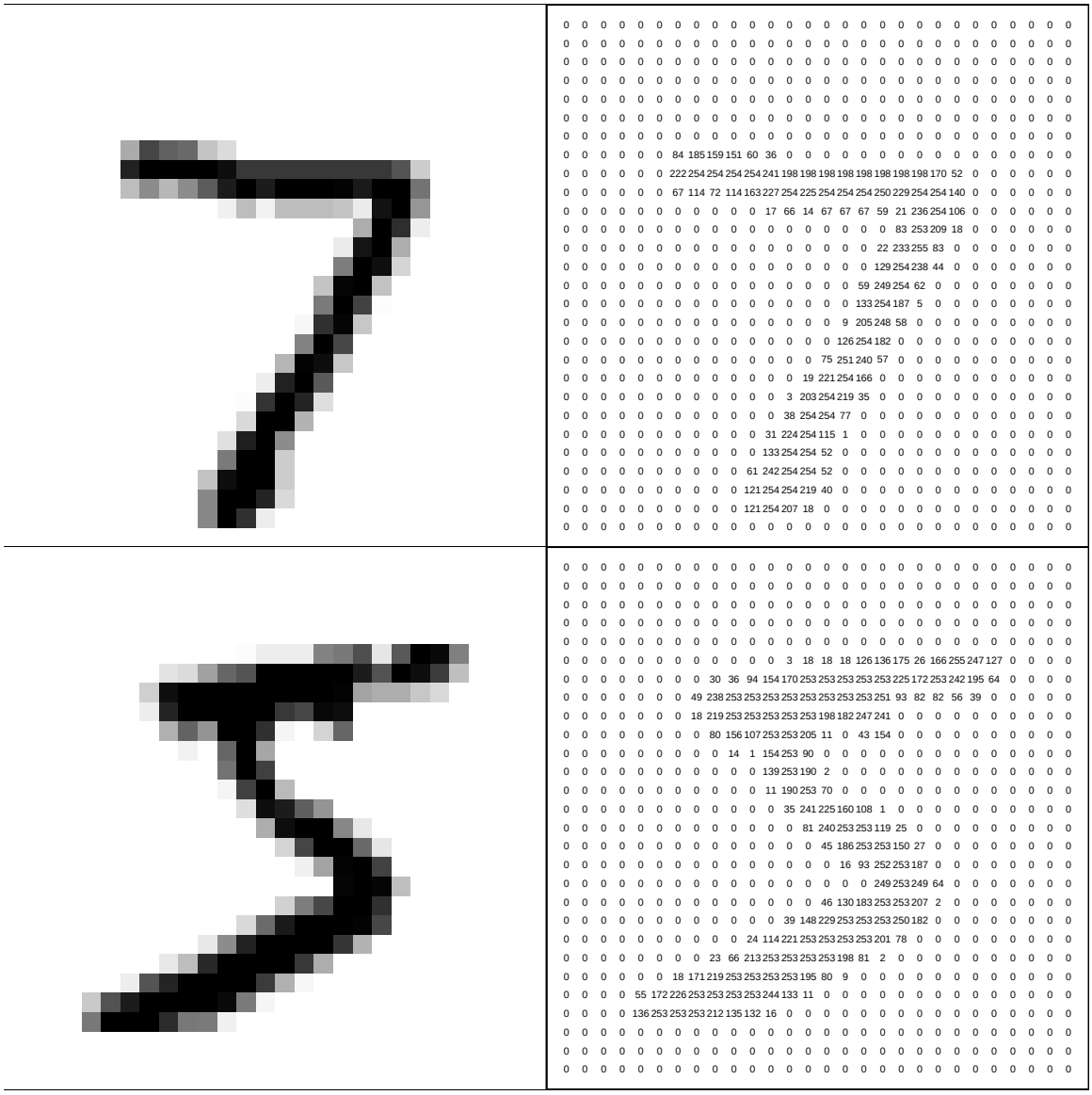


Figure 2.5: Top row: the first test image and its numerical representation $\tilde{\mathbf{x}}_1$. Bottom row: the first image in the training data set and its numerical representation $\mathbf{x}_1$. To compute the similarity between the images, we square the differences between the numbers at the same corresponding positions in the right hand column, and sum.

We compute $d(\tilde{\mathbf{x}}_1, \mathbf{x}_i)$ for all 60000 images. The image most similar to $\tilde{\mathbf{x}}_1$ turns out to be image number 53844 in the training data set:

$$d(\tilde{\mathbf{x}}, \mathbf{x}_{53844}) = \sum_{j=1}^{784} (\tilde{x}_{1,j} - x_{53844,j})^2 \tag{2.3}$$

$$= 457,766 \tag{2.4}$$

We view this below.

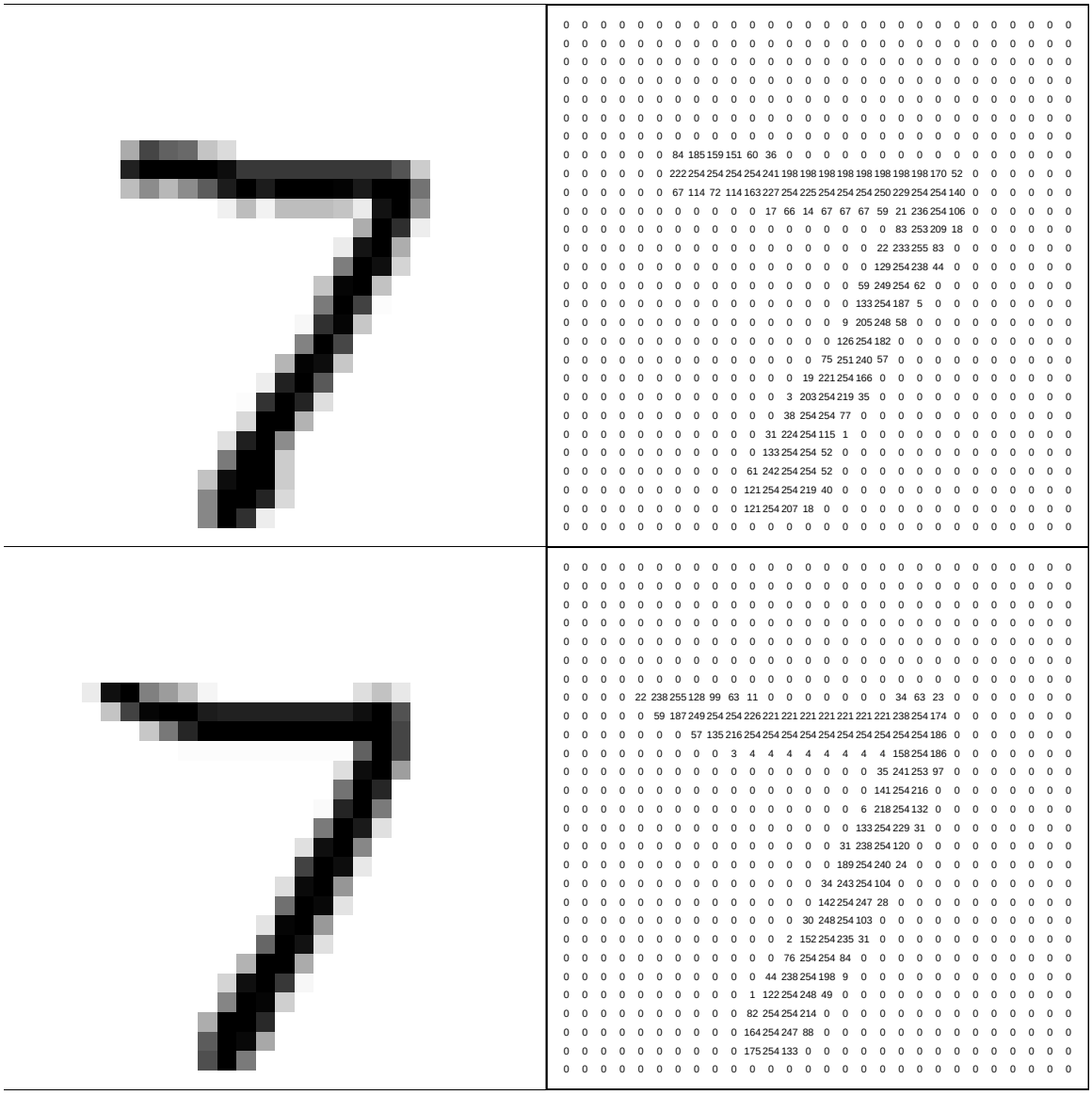


Figure 2.6: Top row: the first test image and its numerical representation $\tilde{\mathbf{x}}_1$. Bottom row: image number 53844 in the training data set and its numerical representation $\mathbf{x}_{53844}$. To compute the similarity between the images, we square the differences between the numbers at the same corresponding positions in the right hand column, and sum. By this measure of similarity, image 53844 is closest to our test image.

So, the computer will look up the value of y_{53844} in the training data set, which is recorded as 7, and so estimate the test image to be the number 7.

2.1.4.1 The 'K' in K nearest neighbours

We've actually used the simplest version of the KNN algorithm, where we look for the single nearest neighbour. An extension is to look for the K nearest neighbours, and then choose the class label based on which which label occurs the most out of the K nearest neighbours (so we have just used $K = 1$ above). This may give better results, for example, if there is the odd 'badly drawn' image in

the training data set, that looks like a different digit: it can be out-voted if we search for more nearest neighbours.

2.1.5 Using K nearest neighbours in R

We can use the function `knn()` from the package `class`.

2.1.5.1 A simple example

We'll first do a simple example on a small data set, to make it easier to see how everything works.

(Ignore the following three commands, unless you want to try this on your own computer. These commands will make the data we are going to use for the example.)

```
irisTrain <- iris[c(1, 6, 51, 52, 101, 102), 3:5]
irisTest <- cbind(flower = c("A", "B"),
                  iris[c(44, 53), 3:4])
row.names(irisTrain) <- row.names(irisTest) <- NULL
```

In our training data, a data frame called `irisTrain`, we have observations of the lengths and widths of petals for 6 flowers (species of iris). Each flower is one of three possible species: setosa, versicolor, or virginica.

```
irisTrain
```

```
##   Petal.Length Petal.Width    Species
## 1          1.4         0.2     setosa
## 2          1.7         0.4     setosa
## 3          4.7         1.4 versicolor
## 4          4.5         1.5 versicolor
## 5          6.0         2.5  virginica
## 6          5.1         1.9  virginica
```

In our test data, `irisTest`, we have two iris flowers, labelled `A` and `B`, with measured petal lengths and widths: the aim is to predict the species of these two flowers.

```
irisTest
```

```
##   flower Petal.Length Petal.Width
## 1      A          1.6         0.6
## 2      B          4.9         1.5
```

If we plot the training and test data together, we can see that the closest flower in the training data to flower A has species setosa, and the closet flower in the training data to flower A has species versicolor, so we would predict that flowers A and B are species setosa and versicolor respectively.

```
ggplot(irisTrain, aes(x = Petal.Length,
                      y = Petal.Width))+
  geom_point(aes(color = Species)) +
  annotate("text", x = irisTest$Petal.Length,
           y = irisTest$Petal.Width,
           label = irisTest$flower)
```

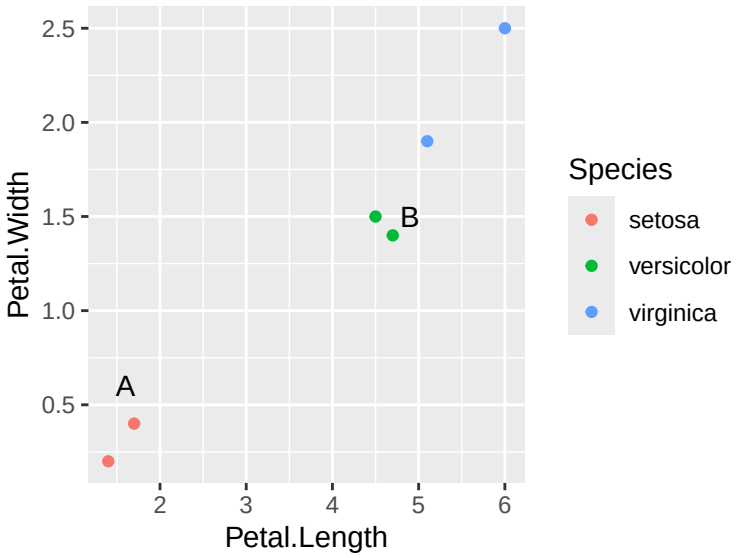


Figure 2.7: The training data are shown as the 6 coloured points. The test data are marked as the two letters. The nearest neighbour to flower A has species setosa, and the nearest neighbour to flower B has species versicolor.

To use the KNN algorithm in R, we use the function `knn()` from the `class` library. We specify three arguments:

- `train`: the measurements in the training data, **excluding** the class labels (the `Species` column). We need to exclude column 3, which we can do as follows:

```
irisTrain[, -3]
```

```
##   Petal.Length Petal.Width
## 1          1.4         0.2
## 2          1.7         0.4
## 3          4.7         1.4
## 4          4.5         1.5
## 5          6.0         2.5
## 6          5.1         1.9
```

- `test`: the measurements in the test data. We will need to exclude the `flower` labels. These labels are also in the first column, so we do

```
irisTest[, -1]
```

```
##   Petal.Length Petal.Width
## 1          1.6         0.6
## 2          4.9         1.5
```

- `cl`: the class labels in the training data. These are in the `Species` column, so we can extract these using

```
irisTrain$Species
```

```
## [1] setosa     setosa     versicolor versicolor virginica  virginica
## Levels: setosa versicolor virginica
```

So, we use the `knn()` function as follows:

```
library(class)
knn(train = irisTrain[, -3],
    test = irisTest[, -1],
    cl = irisTrain$Species)
```

```
## [1] setosa     versicolor
## Levels: setosa versicolor virginica
```

So, out of the three possible `Levels`, the KNN algorithm classifies flower `A` as `setosa` and flower `B` as versicolor. This agrees with what we could see in Figure 2.7

2.1.5.2 Using `knn()` with the handwritten digits

Returning to the handwritten digits, suppose we have the training images and class labels stored in a single data frame called `training_set`, where each row is one image, with the class label in column 1 and the pixel values in columns 2 to 785. We display the first 5 columns below:

```
training_set[, 1:5]
```

```
## # A tibble: 60,000 x 5
##       X1    X2    X3    X4    X5
##    <dbl> <dbl> <dbl> <dbl> <dbl>
##  1     5     0     0     0     0
##  2     0     0     0     0     0
##  3     4     0     0     0     0
##  4     1     0     0     0     0
##  5     9     0     0     0     0
##  6     2     0     0     0     0
##  7     1     0     0     0     0
##  8     3     0     0     0     0
##  9     1     0     0     0     0
## 10     4     0     0     0     0
## # i 59,990 more rows
```

(Look at the first column, and compare it with the first row in Figure 2.1, which shows the first 10 images in the training data set.)

We now extract the class labels and images as follows:

```
training_images <- training_set %>%
  select(-X1)
```

The minus sign means that we select all columns *apart* from `X1`. We then extract the class labels:

```
training_labels <- training_set$X1
```

Suppose the test set are arranged in another data frame `test_set`, with the same structure. We then extract the images:

```
test_images <- test_set %>%
  select(-X1)
```

Now we can use the `knn()` function. We'll just use it on the first 5 images in the test set:

```
library(class)
knn(train = training_images,
    test = test_images[1:5, ],
    cl = training_labels)
```

```
## [1] 7 2 1 0 4
## Levels: 0 1 2 3 4 5 6 7 8 9
```

(The first row of the output means that the algorithm estimates the first five digits to be 7, 2, 1, 0, 4. The second row gives the full range of digits provided in `training_labels`.) Inspecting the first five test images, we can see that the algorithm has worked!

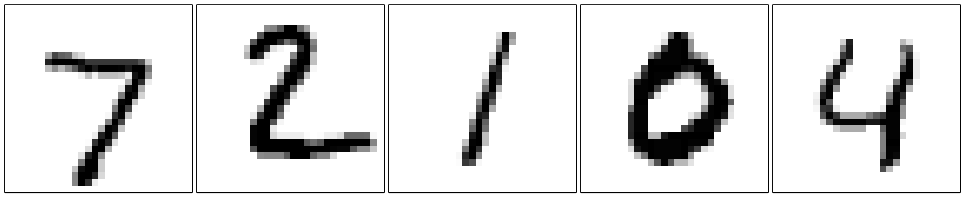


Figure 2.8: The first five images in the test set. We can see that algorithm has correctly identified all five digits.

2.1.6 The performance of the algorithm

The algorithm won't always get it right! Here's an example where the algorithm gets it wrong (test image number 116):

```
knn(train = training_images,
    test = test_images[116, ],
    cl = training_labels)
```

```
## [1] 9
## Levels: 0 1 2 3 4 5 6 7 8 9
```

After a little investigation (details omitted), it turns out that image 8112 in the training data is closest to image 116 in the test data. If we look at the image, we can see where the algorithm went wrong: the test image is a '4', but it looks very similar to a '9' in the training data.

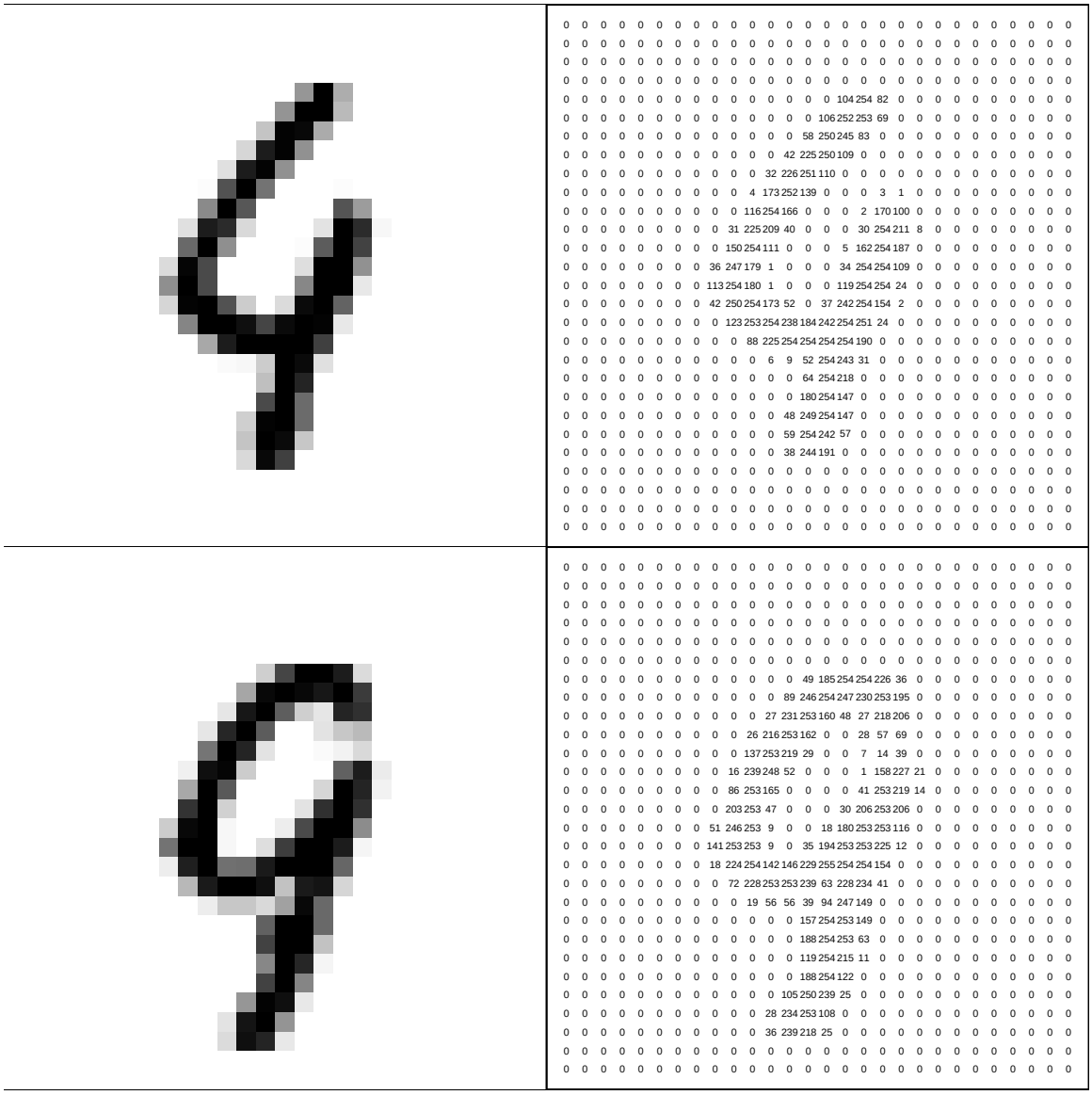


Figure 2.9: Top row: test image 116 and its numerical representation $\tilde{\mathbf{x}}_{116}$. Bottom row: image number 8112 in the training data set and its numerical representation $\mathbf{x}_{8112}$. This training image is the closest to our test image, but the digits are not the same!

Applying the algorithm to all 10000 test images, the algorithm gets the right answer 9691 times (96.91%). This may look quite good, but an error rate of 3% is probably too large in practice. But the performance is good enough to show the potential of using a data-based method for the image recognition problem. In fact, more complex methods do give better performance. A list of results is maintained here, with the best performing method (at the time of writing) being 99.79% accurate.

Chapter 3

Populations, samples and statistical models

We describe the problem of inferring characteristics of a **population**, such as a population mean or proportion, given a random **sample** drawn from that population. One general approach for learning about populations from samples is to model the sample observations as random draws from a probability distribution, with the probability distribution representing the whole population. The parameters of the distribution can then be interpreted as describing population characteristics, such as the mean of the population.

3.1 Statistical models

Given a general description of a problem, we have to decide what probability distribution to use a suitable model, and think how the parameter(s) of that distribution relates to our quantity of interest

Example 3.1 (Choosing probability distributions to represent data.)**.** Here are four examples.

1. In a crime survey, we wish to estimate the proportion of households in a city that have been burgled in the last year. One hundred households are to be selected at random. Define X to be the number of households responding that they have been burgled. Choose a suitable probability distribution for X, and relate the parameter(s) in that distribution to the quantity of interest.

Solution.

X is a discrete random variable. If we treat the 100 (randomly selected) households as independent, we can think of X as the number of times an event happens out of 100 trials. We model this with a binomial distribution:

$$X \sim Bin(100, \theta).$$

The parameter θ would be the probability of a single household responding that they have been burgled; we interpret θ as the population proportion of all burgled households.

It is unlikely that the sample proportion X/n will equal the true (but unknown) population proportion θ. By studying this model, we can understand how X/n may deviate from θ (more in the next chapter).

2. The number of accidents occurring over a road network is to be investigated. The interest is in the mean number of accidents per week, averaged over 'all' weeks. Define $X_1, \ldots, X_{10}$ to be the number of accidents that will be observed in each of the first 10 weeks. Choose a suitable probability distribution for $X_1, \ldots, X_{10}$, and relate the parameter(s) in that distribution to the quantity of interest.

Solution.

Each X_i is a discrete random variable, and is a count of the number of accidents in a particular week. We are counting the number of events in a period of time: this situation can be modelled with a Poisson distribution:

$$X_1, \ldots, X_{10} \overset{i.i.d}{\sim} Poisson(\lambda).$$

The notation i.i.d is short for "independent and identically distributed". Each of $X_1, \ldots, X_{10}$ has the same $Poisson(\lambda)$ distribution, and $X_1, \ldots, X_{10}$ are assumed independent.

We interpret λ as the population mean number of accidents per week, over all weeks, assuming no change to the underlying risk of an accident occurring in any week.

3. In a call-centre, there is interest in the volume of incoming calls during peak opening times. Specifically, there is interest in the mean length of time (in seconds) between two successive incoming calls. Define $X_1, \ldots, X_{100}$ to be the times that will be observed between calls for 100 successive pairs of calls. Choose a suitable probability distribution for $X_1, \ldots, X_{100}$, and relate the parameter(s) in that distribution to the quantity of interest.

Solution.

The time between two successive calls is a continuous quantity, cannot be negative, but could be close to 0. From the distributions you have met so far in this module, the exponential distribution would be the most suitable choice. We suppose

$$X_1, \ldots, X_{100} \overset{i.i.d}{\sim} Exponential(\lambda),$$

and we interpret $1/\lambda$ as the population mean time between all successive calls during peak hours.

4. For an electric car, we are interested in how many miles the car can be driven on a single charge of the battery, on a particular test journey route. The interest is in both what the mileage is 'on average', and how variable this mileage can be. Eight cars, all of the same model, are to be driven on the test route, all starting at the same time and day each week. Let $X_1, \ldots, X_8$ denote the mileages that will be observed. Choose a suitable probability distribution for $X_1, \ldots, X_8$, and relate the parameter(s) in that distribution to the quantities of interest.

Solution.

Mileage is a continuous quantity. Out of the continuous distributions we have covered in this module, the most suitable one would be a normal distribution:

$$X_1, \ldots, X_8 \overset{i.i.d}{\sim} N(\mu, \sigma^2).$$

We interpret μ as the population mean mileage, and σ^2 as the population variance of the mileages,

for all cars of this model on the test route.

3.1.1 Objectives

We have described various scenarios where we are interested in some characteristic of a population (mean, variance, proportion), and we draw a sample from that population. In the remaining chapters, we will study

- how to estimate these population characteristics, once we have observed our data, via **estimating parameters of probability distributions**;
- how to assess the accuracy of our estimates, using **confidence intervals**;
- how to compare different populations, and test if they have different characteristics or not, using a variety of **hypothesis tests**.

3.1.2 Comment: infinite and finite populations

When a population is represented by a probability distribution, we are working with what we describe as an **infinite** population: each member of the population corresponds to a random draw from this distribution, and there is no limit on the number of possible random draws: there is no limit on the size of the population.

An alternative is the **finite population** approach, in which we define a (finite) list of all population members, and any sample of data is a random selection from this list. With finite population methods, we don't actually need to assume a probability distribution for the population members: we can do inference for population means, proportions etc. without assuming one. The limitation with this is that there are many statistical inference problems where the sample can't be thought of as a random selection from a list. For example, if the population of interest was all newborn babies, including babies born in the future, then some population members do not yet exist; we could not produce a list of all the population members.

In this module, we will cover infinite population approaches only. Finite population methods are covered in MPS318: Sampling Theory and Design of Experiments.

Chapter 4

Point estimation

In this chapter, we consider how to estimate the parameters of a probability distribution, given observations from that distribution. By "point" estimate, we mean a single number as the estimate. (In the next chapter, we will consider *interval* estimates: estimates in the form of a *range* of values).

4.1 Estimating the parameters of a normal distribution

4.1.1 Problem setup and notation

Suppose we model some data with a normal distribution: we have n independent and identically distributed normal random variables

$$X_1, X_2, \dots, X_n \overset{i.i.d}{\sim} N(\mu, \sigma^2), \tag{4.1}$$

where the values of μ and σ^2 are unknown to us. We denote the *observed values* of these n random variables by $x_1, \dots, x_n$. How should we estimate μ and σ^2 using $x_1, \dots, x_n$? This problem is illustrated in Figure 4.1.

> The distinction between big X and little x is important: we use X_i to represent a **random variable**, and x_i to represent the **observed value** of a random variable.

- We can think of X_i as describing the situation *before* we have collected our data: we don't know what value we are going to observe, so we represent it by a random variable X_i.
- We can think of x_i as describing the situation *after* we have collected our data: we now have a numerical value for our ith observation, which we denote by a constant x_i.

4.1.2 The sample mean and sample variance

We'll define some notation/terminology now, that we will use a lot from now on.

Definition 4.1 (Sample mean)**.** Given the observations $x_1, \dots, x_n$, we define the sample mean to be

$$\bar{x} := \frac{1}{n} \sum_{i=1}^{n} x_i.$$

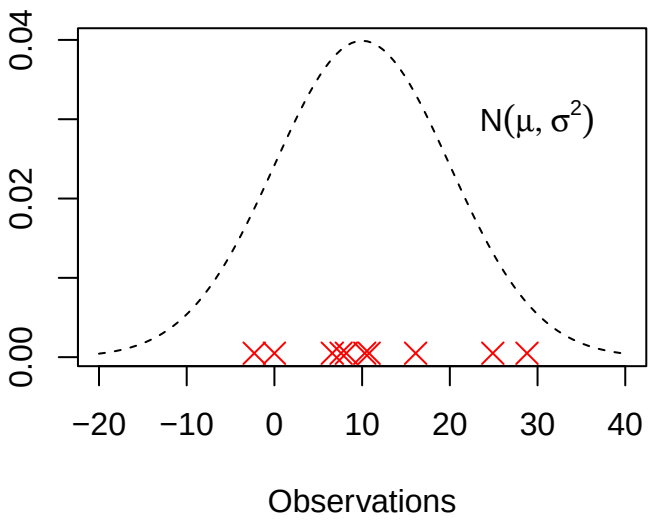


Figure 4.1: The red crosses show 10 observations drawn from the $N(\mu, \sigma^2)$ distribution. Suppose we *only* know the values of these 10 observations: can we use them to estimate the values of μ and σ^2?

Definition 4.2 (Sample variance)**.** Given the observations $x_1, \ldots, x_n$, we define the sample variance to be

$$s^2 := \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2.$$

Confusingly, the word "mean" has multiple meanings in probability and statistics! Don't confuse the "sample mean" $\bar{x}$ with the "mean of a random variable" $\mathbb{E}(X)$: they are not the same thing! Similarly, "sample variance" doesn't mean the same thing as the "variance of a random variable" $Var(X)$

4.1.3 Point estimates for the mean and variance

For now, we will simply state suitable estimates for the parameters of any distribution (methods such as "maximum likelihood estimation" can be used to derive estimates). For the normal distribution, we will use the estimates

$$\hat{\mu} := \bar{x} \tag{4.2}$$

$$\hat{\sigma}^2 := s^2. \tag{4.3}$$

The hat $\hat{}$ notation is important here: it is used to denote that $\hat{\mu}$ and $\hat{\sigma}^2$ are only *estimates* of μ and σ^2. Don't write an expression such as $\mu = \bar{x}$: this would be claiming that the *true* value of the unknown mean parameter μ is the sample mean $\bar{x}$.

In R, these can be calculated with the functions `mean()` and `var()` respectively.

4.1.4 Testing the method

To illustrate the notation, the R commands, and to persuade ourselves that the formulae give sensible estimates, we'll do a simulation example in R.

We'll consider a sample of 100 observations

$$X_1, \ldots, X_{100} \overset{i.i.d}{\sim} N(\mu, \sigma^2)$$

We'll now get R to generate some data. To do this, we have to choose values for μ and σ^2: I'll choose $\mu = 20$ and $\sigma^2 = 25$:

```
x <- rnorm(n = 100, mean = 20, sd = 5)
```

If we inspect the first three elements of `x`, we get

```
x[1:3]
```

```
## [1] 20.36 16.76 16.07
```

so we have $x_1 = 20.36$, $x_2 = 16.76$, $x_3 = 16.07$, ... Now pretending that we don't know the true values of μ and σ^2, we compute our sample mean and sample variance:

```
mean(x)
```

```
## [1] 18.89395
```

```
var(x)
```

```
## [1] 30.37141
```

so we have

$$\begin{aligned} \hat{\mu} &= \bar{x} = \frac{1}{100}\sum_{i=1}^{100} x_i = 18.9, \\ \hat{\sigma}^2 &= s^2 = \frac{1}{99}\sum_{i=1}^{100}(x_i - \bar{x})^2 = 30.4, \end{aligned}$$

to 3 s.f., and these estimates are similar to the true values, as we would hope. We illustrate this in Figure 4.2.

4.1.5 Estimators and estimates

Over the next few sections, we will explain why $\bar{x}$ and s^2 were good choices of estimates to use for μ and σ^2.

As a motivating example, suppose we are planning an experiment to test a new steel manufacturing process. A batch of n steel cables will be produced, and the tensile strength of each cable will be measured. We don't yet know what the tensile strengths will be, and we denote the strength of the ith cable by the random variable X_i. We suppose that the strengths are normally distributed:

$$X_1, X_2, \ldots, X_n \overset{i.i.d}{\sim} N(\mu, \sigma^2), \tag{4.4}$$

and the aim of the experiment is to estimate μ and σ^2.

> Although we don't yet have the data, we are going to declare **now** how we intend to estimate μ and σ^2, and then investigate whether our estimation method is likely to 'work' or not: whether it is likely to produce good estimates.

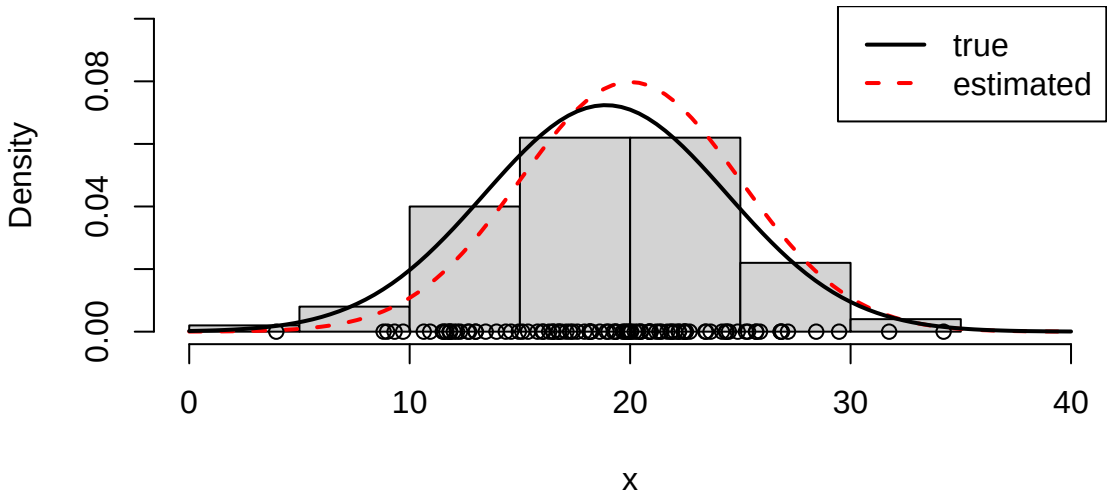


Figure 4.2: The circles and histogram represent 100 observations drawn from the $N(\mu, \sigma^2)$ distribution. We display the true normal distribution with the red dashed line. We compute estimates of μ and σ^2 using $\bar{x}$ and s^2: the corresponding estimated normal distribution is shown as the black curve, and is fairly similar to the true distribution: the estimates are fairly similar to the true values.

We declare that we are going to use the following two **estimators**, to estimate μ and σ^2 respectively:

$$\bar{X} := \frac{1}{N}\sum_{i=1}^{n} X_i, \tag{4.5}$$

$$S^2 := \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2. \tag{4.6}$$

Definition 4.3 (Estimators)**.** An estimator is a function of some random variables $X_1, \ldots, X_n$ intended to provide an estimate of some parameter from the distribution of $X_1, \ldots, X_n$.

Don't confuse $\bar{X}$ and S^2 with $\bar{x}$ and s^2! Both $\bar{X}$ and S^2 are random variables because they are functions of the random variables $X_1, \ldots, X_n$.

An estimat**or** is a random variable; the observed value of the estimator, calculated using the observed values $x_1, \ldots, x_n$ is the corresponding estimat**e**.

In our steel cables example:

- we haven't yet done the experiment; the cables haven't been manufactured yet, and we can't know what values for the tensile strengths we will observe. At this stage, we think of the n tensile strengths $X_1, \ldots, X_n$ as random variables, so both $\bar{X}$ and S^2 are random too.
- After we have done the experiment, we will have the n numerical values for the tensile strengths, and the sample mean and variance that we calculate from these values will be represented by the constants $\bar{x}$ and s^2.

Now we can ask the question: are $\bar{X}$ and S^2 good estimators? Given that they are random variables, how likely is it that they will give us values 'close' to the true values of μ and σ^2?

4.1.6 The χ^2 distribution

We'll shortly discuss the distribution of the two estimators $\bar{X}$ and S^2, but first, we need to introduce a new distribution.

Definition 4.4 (χ^2 distribution)**.** If a random variable Y has the χ^2_ν distribution (the "chi-squared distribution with ν degrees of freedom"), we write $Y \sim \chi^2_\nu$ and the probability density function of Y is given by

$$f_Y(y) = \begin{cases} \frac{y^{\nu/2-1}}{2^{\nu/2}\Gamma(\nu/2)} \exp\left(-\frac{y}{2}\right), & y \geq 0, \\ 0 & y < 0. \end{cases} \tag{4.7}$$

Note that we must have $\nu > 0$. Here Γ denotes the gamma function, defined by

$$\Gamma(x) = \int_0^\infty t^{x-1} e^{-t} dt. \tag{4.8}$$

It can be shown that

$$\mathbb{E}(Y) = \nu, \tag{4.9}$$

$$Var(Y) = 2\nu. \tag{4.10}$$

The χ^2_ν distribution is positively skewed, with the skew more apparent as the degrees of freedom ν decreases. Three χ^2 distributions are plotted in Figure 4.3.

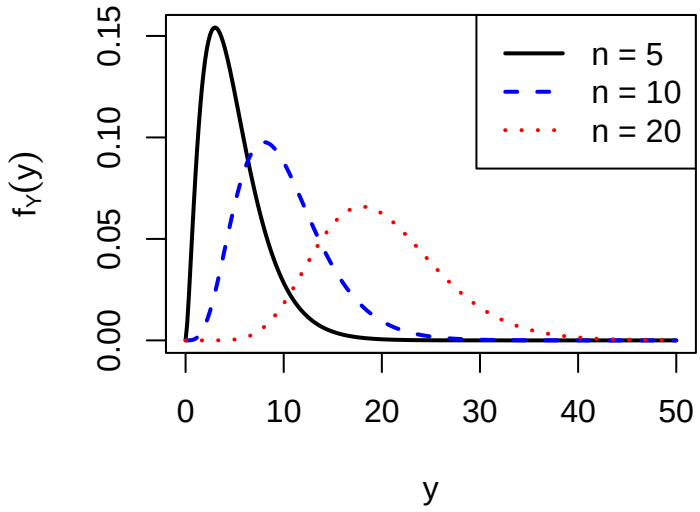


Figure 4.3: Three χ^2_n distributions with $n = 5$, 10 and 20 degrees of freedom.

We won't be working with the density function of the χ^2 distribution in this module, but you will need to know its mean and variance, and that a χ^2 random variable cannot be negative.

4.1.7 The distribution of the estimators

As the estimators are random variables, we can derive their probability distributions. It can be shown that:

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right) \tag{4.11}$$

For the distribution of S^2, we have the following result. If define the random variable R to be

$$R := \frac{(n-1)S^2}{\sigma^2}, \tag{4.12}$$

then R has the probability distribution

$$R \sim \chi^2_{n-1}. \tag{4.13}$$

You can learn how to prove this result in MPS235. For now, we'll just note that $\frac{(n-1)S^2}{\sigma^2}$ cannot be negative, which is one property of the χ^2 distribution.

4.1.8 Unbiased estimators

We can now justify why $\bar{X}$ and S^2 are good choices of estimators for μ and σ^2.

Definition 4.5 (Unbiased estimator)**.** Let $T(X_1, \dots, X_n)$ be a function of $X_1, \dots, X_n$. We say that $T(X_1, \dots, X_n)$ is an unbiased estimator of a parameter θ if

$$\mathbb{E}(T(X_1, \dots, X_n)) = \theta$$

Informally, we could say that an unbiased estimator is 'expected to give the right answer', or 'gives the right answer on average'.

$\bar{X}$ and S^2 are unbiased estimators of μ and σ^2 respectively: we have

$$\mathbb{E}(\bar{X}) = \mu, \tag{4.14}$$
$$\mathbb{E}(S^2) = \sigma^2. \tag{4.15}$$

Example 4.1 (Unbiased estimators: sample mean and sample variance)**.** Prove that $\bar{X}$ and S^2 are unbiased estimators of μ and σ^2.

Solution

Recall the basic properties of expectations:

$$\mathbb{E}(aX) = a\mathbb{E}(X), \tag{4.16}$$
$$\mathbb{E}(X_i + X_j) = \mathbb{E}(X_i) + \mathbb{E}(X_j). \tag{4.17}$$

Now, for $\bar{X}$ to be an unbiased estimator of μ, we need to show that $\mathbb{E}(\bar{X}) = \mu$. We have

$$\mathbb{E}(\bar{X}) = \mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) \tag{4.18}$$

$$= \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}(X_i) \tag{4.19}$$

$$= \frac{1}{n}\sum_{i=1}^{n} \mu \tag{4.20}$$

$$= \frac{n\mu}{n} \tag{4.21}$$

$$= \mu, \tag{4.22}$$

as required.

For S^2 to be an unbiased estimator of σ^2, we need to show that $\mathbb{E}(S^2) = \sigma^2$. Recall that we defined $R = \frac{(n-1)S^2}{\sigma^2}$, and stated that $R \sim \chi^2_{n-1}$, which means that $\mathbb{E}(R) = n - 1$, using the result in (4.9)

We have

$$\mathbb{E}(S^2) = \mathbb{E}\left(\frac{R\sigma^2}{n-1}\right) \tag{4.23}$$

$$= \frac{\sigma^2}{n-1}\mathbb{E}(R) \tag{4.24}$$

$$= \frac{\sigma^2}{n-1} \times (n-1) \tag{4.25}$$

$$= \sigma^2, \tag{4.26}$$

4.1.9 The standard error of an estimator

Definition 4.6 (Standard error)**.** Let $T(X_1, \ldots, X_n)$ be a function of $X_1, \ldots, X_n$, used to estimate some parameter θ. The standard error of $T(X_1, \ldots, X_n)$ is defined to be

$$s.e.(T(X_1, \ldots, X_n)) = \tag{4.27}$$

$$\sqrt{Var(T(X_1, \ldots, X_n))}, \tag{4.28}$$

i.e the square root of its variance.

As the standard error is the square root of a variance, we could also call it a 'standard deviation'. However, when referring to estimators, we use the term 'standard error' instead.

For an unbiased estimator, the smaller the standard error the better: the closer the estimate is likely to be to the true value.

Writing $T = T(X_1, \ldots, X_n)$ for short, for an unbiased estimator T of a parameter θ, we have

$$Var(T) := \mathbb{E}((T - \mathbb{E}(T))^2) = \mathbb{E}((T - \theta)^2)$$

so the smaller the standard error, the smaller the expectation of the (squared) difference between T and θ.

We have

$$s.e.(\bar{X}) = \sqrt{\frac{\sigma^2}{n}}, \tag{4.29}$$

$$s.e.(S^2) = \sqrt{\frac{2\sigma^4}{n-1}}. \tag{4.30}$$

Example 4.2 (Standard error of the sample mean)**.** Derive the standard error of $\bar{X}$

Solution

Recall the basic properties of variances:

$$Var(aX) = a^2 Var(X),$$

and for X_i and X_j independent, we have

$$Var(X_i + X_j) = Var(X_i) + Var(X_j).$$

(if they are not independent, we have to include the covariance term $Var(X_i + X_j) = Var(X_i) + Var(X_j) + 2Cov(X_i, X_j)$)

We have

$$Var(\bar{X}) = Var\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) \tag{4.31}$$

$$= \frac{1}{n^2}\sum_{i=1}^{n} Var(X_i) \tag{4.32}$$

$$= \frac{1}{n^2}\sum_{i=1}^{n} \sigma^2 \tag{4.33}$$

$$= \frac{n\sigma^2}{n^2} \tag{4.34}$$

$$= \frac{\sigma^2}{n}, \tag{4.35}$$

so we have $s.e.(\bar{X}) = \sqrt{\sigma^2/n}$.

Example 4.3 (Standard error of the sample variance)**.** Derive the standard error of S^2.

Solution

Recall that we defined $R = \frac{(n-1)S^2}{\sigma^2}$, and stated that $R \sim \chi^2_{n-1}$, which means that $Var(R) = 2(n-1)$, using the result in (4.10)

We have

$$Var(S^2) = Var\left(\frac{R\sigma^2}{n-1}\right) \tag{4.36}$$

$$= \frac{\sigma^4}{(n-1)^2} Var(R) \tag{4.37}$$

$$= \frac{\sigma^4}{(n-1)^2} \times 2(n-1), \tag{4.38}$$

and so $s.e.(S^2) = \sqrt{\frac{2\sigma^4}{n-1}}$.

4.1.10 Consistent estimators

Definition 4.7 (Consistent estimator)**.** An estimator $T(X_1, \ldots, X_n)$ for a parameter θ is consistent if, for any $\epsilon > 0$, we have

$$\lim_{n\to\infty} P(|T(X_1, \ldots, X_n) - \theta| < \epsilon) = 1. \tag{4.39}$$

Informally, we can say that a consistent estimator is guaranteed to converge to the true value of the parameter as the sample size tends to infinity.

Theorem 4.1 (Identifying a consistent estimator)**.** *If an estimator is unbiased, and its standard error tends to 0 as the sample size n tends to infinity, it will also be a consistent estimator.*

You do not need to know the proof of this result for this module, but if you want a challenge, you are encouraged to prove this result for yourself! (Hint: you will need to use Chebyshev's inequality.) A proof is provided in the tutorial booklet solutions.

Both $\bar{X}$ and S^2 are consistent estimators, which gives another justification for using them.

Example 4.4 (Consistency of sample mean and sample variance)**.** Verify that $\bar{X}$ and S^2 are consistent estimators for μ and σ^2.

Solution

We have already shown, in Example 4.1 that $\bar{X}$ and S^2 are unbiased estimator for μ and σ^2 respectively. We have also shown, in Examples 4.2 and 4.3 that the standard errors are

$$s.e.(\bar{X}) = \sqrt{\frac{\sigma^2}{n}}, \tag{4.40}$$

$$s.e.(S^2) = \sqrt{\frac{2\sigma^4}{n-1}}. \tag{4.41}$$

We can see that both these standard errors tend to 0 as $n \to \infty$, so both estimators are consistent.

4.2 Estimating the probability parameter in a Binomial distribution

Suppose we have a single random variable

$$X \sim Bin(n, \theta)$$

with x the observed value of X: the observed number of 'successes' in n trials. If θ is unknown, how should we estimate it? (The other parameter in the distribution, n, would typically be known).

Here, the obvious choice would be

$$\hat{\theta} = \frac{x}{n},$$

so our estimator is $\frac{X}{n}$. It can be shown that this is an unbiased estimator of θ, is consistent, and has standard error $\sqrt{\frac{\theta(1-\theta)}{n}}$.

Example 4.5 (Unbiased estimators: sample proportion)**.** Prove that $\frac{X}{n}$ is unbiased estimator of θ for $X \sim Bin(n, \theta)$.

Solution

For $X \sim Bin(n, \theta)$, we have $\mathbb{E}(X) = n\theta$. To show that X/n is an unbiased estimator of θ, we need to show $\mathbb{E}(X/n) = \theta$. We have

$$\mathbb{E}\left(\frac{X}{n}\right) = \frac{1}{n}\mathbb{E}(X) \tag{4.42}$$

$$= \frac{n\theta}{n} \tag{4.43}$$

$$= \theta, \tag{4.44}$$

as required.

Example 4.6 (Standard error of the sample proportion)**.** Derive the standard error of $\frac{X}{n}$ for $X \sim Bin(n, \theta)$

Solution

For $X \sim Bin(n, \theta)$, we have $Var(X) = n\theta(1-\theta)$. For the standard error, we have

$$Var\left(\frac{X}{n}\right) = \frac{1}{n^2}Var(X) \tag{4.45}$$

$$= \frac{n\theta(1-\theta)}{n^2}, \tag{4.46}$$

and so we have

$$s.e.(X/n) = \sqrt{\frac{\theta(1-\theta)}{n}}.$$

Example 4.7 (Consistency of sample proportion)**.** Verify that $\frac{X}{n}$ is a consistent estimator of θ for $X \sim Bin(n, \theta)$.

Solution.

We have shown that $\frac{X}{n}$ is an unbiased estimator of θ. To be consistent, we require its standard error to tend to 0 as $n \to \infty$. We have shown that the standard error is $\sqrt{\theta(1-\theta)/n}$, so the requirement is met: $\frac{X}{n}$ is a consistent estimator for θ.

Chapter 5

Interval estimates and confidence intervals

In the last chapter we learnt how to obtain *point* estimates for parameters in probability distributions. The problem with point estimates is that they will almost certainly be *wrong*! Sample means will almost always differ from population means, for example.

Definition 5.1 (Interval estimate)**.** We use the term "interval estimate" to mean a range of values that we think are plausible for some unknown parameter. For example, instead of reporting a point estimate: "we estimate μ to be 11.5", we might report an interval estimate: "we think μ is between 9.5 and 13.5".

> By providing an interval estimate, we are able to describe our **uncertainty** about a parameter: the more uncertain we are, the wider the interval.

In this chapter, we will study a particular type of interval estimate known as a *confidence interval*. First, we need a little more distribution theory

5.1 The Student t distribution

We will use the Student t distribution shortly for obtaining confidence intervals.

Definition 5.2 (Student t distribution)**.** If a random variable Y has a Student t distribution (or "Student's t" distribution or just "t distribution" for short) with ν degrees of freedom, that is if

$$Y \sim t_\nu,$$

then Y has the density function

$$f_\nu(y) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)\sqrt{\pi\nu}}\left(1+\frac{y^2}{\nu}\right)^{-\frac{\nu+1}{2}},$$

for $\infty < y < \infty$

5.1.1 Mean and variance of the t-distribution

If $\nu > 1$ then

$$E(Y) = 0,$$

and if $\nu > 2$ then

$$V(Y) = \frac{\nu}{\nu - 2}$$

The pdf f_ν is symmetric about zero. For large values of ν, it is very similar to the standard normal density $N(0,1)$. Some t-distributions are plotted in Figure 5.1.

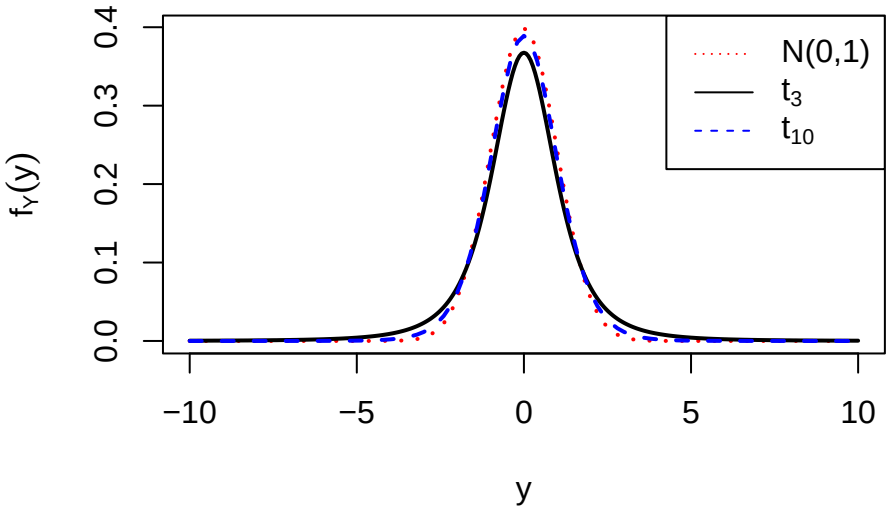


Figure 5.1: The t_3 and t_{10} distributions, together with the standard normal distribution. A t distribution with more than 30 degrees of freedom is hard to distinguish from a standard normal distribution. Note that t distributions have **heavier tails** than the normal.

Theorem 5.1 (Relationship between the normal distribution, the χ^2 distribution and the t distribution)**.** *If $Z \sim N(0,1)$ and $Y \sim \chi^2_\nu$, then*

$$T = \frac{Z}{\sqrt{Y/\nu}} \sim t_\nu,$$

so the ratio of a standard normal variable to the square root of a χ^2 variable has a t distribution.

(We do not prove this result in this module.)

5.1.2 Notation: quantiles/percentiles of the t distribution

If T has a t distribution with ν degrees of freedom, we define $t_{\nu,\ \alpha}$, for $\alpha \in (0,1)$, by

$$P(T \leq t_{\nu,\ \alpha}) = 1 - \alpha,$$

so $t_{\nu,\ \alpha}$ is the $(1-\alpha)$ quantile or $100(1-\alpha)$ percentile of the t_ν distribution. For example, $t_{10;\ 0.05} = 1.812$, so 1.812 is the 95th percentile of the t distribution with 10 degrees of freedom.

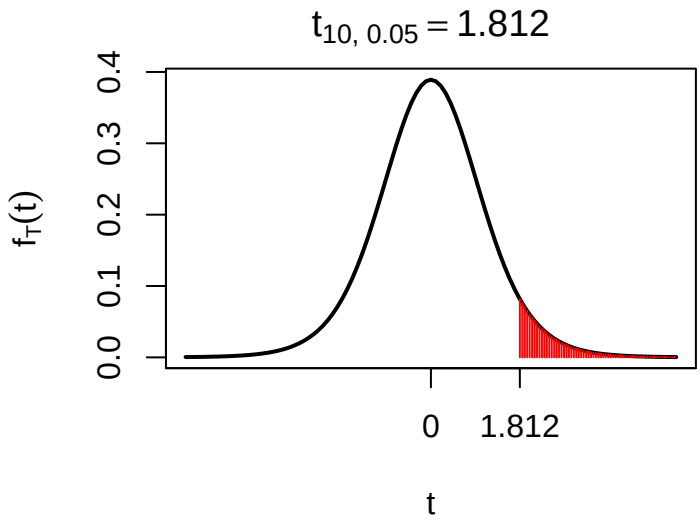


Figure 5.2: The 95th percentile of the t_{10} distribution, which is 1.812 to 3 d.p. Note the convention for the term α in $t_{\nu,\alpha}$ to refer the probability to the right (the shaded area), so that the 95th percentile is denoted by $t_{10,0.05}$.

5.1.3 The t distribution in R.

Cumulative probabilities and quantiles/percentiles can be calculated in R. To calculate a probability, we use the `pt()` command. For example, to calculate $P(T \leq -1)$ for $T \sim t_3$, we do

```
pt(-1, 3)
```

```
## [1] 0.1955011
```

hence, for $T \sim t_3$, we have $P(T \leq -1) = 0.196$ (to 3 d.p.).

To calculate a quantile/percentile, we use the `qt()` command. For example, if we want the 95th percentile of the t_{10} distribution, we do

```
qt(0.95, 10)
```

```
## [1] 1.812461
```

Note that in R, we have specified the left tail probability (0.95), whereas the convention when writing quantiles is to use the right tail probability: we write

$$t_{10,0.05} = 1.812 \text{ to 3 d.p.}$$

5.2 Confidence intervals for the mean and the variance of a normal distribution

Suppose we have n independent and identically distributed normal random variables

$$X_1, X_2, \ldots, X_n \overset{i.i.d}{\sim} N(\mu, \sigma^2),$$

where the values of μ and σ^2 are unknown to us. As usual, denote the *observed values* of these n random variables by $x_1, \ldots, x_n$. We now want to report interval estimates for μ and σ^2, given $x_1, \ldots, x_n$. We will report 95% confidence intervals.

- 95% confidence interval for the mean μ:

$$\bar{x} \pm t_{n-1,0.025}\sqrt{\frac{s^2}{n}}, \tag{5.1}$$

- 95% confidence interval for the variance σ^2:

$$\left[\frac{(n-1)s^2}{\chi^2_{n-1;0.025}}, \frac{(n-1)s^2}{\chi^2_{n-1;0.975}}\right], \tag{5.2}$$

  with

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i, \tag{5.3}$$

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2. \tag{5.4}$$

By inspecting (5.1), we see that

- we'd expect the confidence interval for μ to get narrower as the sample size n increases. The more data we have, the less uncertain we should be.

- a larger s^2 will make the confidence interval wider. A larger s^2 means there is more variability in the data, which makes it harder to get a good estimate of the mean μ.

Although less obvious from (5.2), increasing the sample size should also reduce the width of the interval.

Example 5.1 (Confidence intervals for the mean and variance of a normal distribution: Netflix stock prices)**.** In this example, we will work with some financial data. First, some background. Netflix was one of the best performing stocks in 2018. We will compare it with one other stock: GlaxoSmithKline (GSK). Figure 5.3 (left plot) shows end of day share prices for Netflix and GSK, for each trading day in 2018.

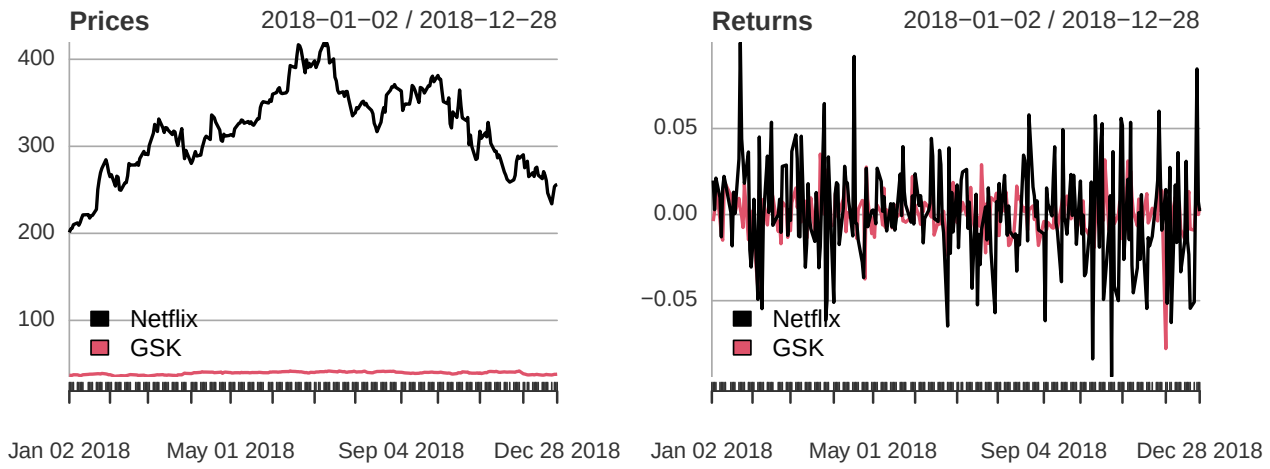


Figure 5.3: Left plot: daily share prices for Netflix and GSK Right plot: daily returns (daily change in share price as a proportion of the price the price at the start of the day). The returns give the profit/loss one would make from one day to the next, and look similar for the two stocks.

For investing, it's not so much the actual price that matters, rather, it's the *return* on the investment that counts. If we define S_i to be the Netflix share price at the end of day i, we define the ith daily return for a Netflix share as

$$X_i = \frac{S_i - S_{i-1}}{S_{i-1}}$$

(e.g., \$1000 invested on day $i-1$ will have grown to $\$(1+X_i)1000$ by the end of day i.) Daily returns are shown in the right plot. We have 249 daily returns, and we suppose

$$X_1, \ldots, X_{249} \overset{i.i.d}{\sim} N(\mu, \sigma^2).$$

We interpret μ as a population mean return. This describes one aspect of the stock's performance: what the *expected* return would be on any given day: if μ turned out to be negative, we would actually expect the stock to decline in value over the long term. The parameter σ is referred to as the **volatility** of the return. Investors care about the volatility as well as the mean return, because it can describe how risky investing in the stock would be.

We'll now state the problem in general terms, without the finance jargon. We have some random variables

$$X_1, \ldots, X_{249} \overset{i.i.d}{\sim} N(\mu, \sigma^2).$$

Given the corresponding observed values $x_1, \ldots, x_{249}$, we want to compute 95% confidence intervals for μ and σ^2.

The observed values $x_1, \ldots, x_{249}$ are stored in R in the vector `netflix`. The first three observations are

```
netflix[1:3]
```

```
## [1] 0.020 0.003 0.021
```

so we have $x_1 = 0.02, x_2 = 0.003, x_3 = 0.021$ and so on.

Task: using the following R output, compute 95% confidence intervals for μ and σ.

```
c(mean(netflix), var(netflix))
```

```
## [1] 0.0014257 0.0008466
```

```
qt(0.975, 248)
```

```
## [1] 1.97
```

```
qchisq(c(0.025, 0.975), 248)
```

```
## [1] 206.3 293.5
```

Solution

First, we consider the R output for the t and χ^2 distributions. We have

which means that $t_{248,\,0.025} = 1.97$ (to 3 d.p.): the 97.5th percentile of the t_{248} distribution is 1.97 (very similar to the normal distribution). This is displayed below: the red shaded region indicates a 2.5% probability of exceeding 1.97.

We also have

which means that $\chi^2_{248,\,0.975} = 206.3$ and $\chi^2_{248,\,0.025} = 293.5$ (to 1 d.p.): the 2.5th and 97.5th percentiles of the χ^2_{248} distribution are 206.3 and 293.5 respectively. These are displayed below: each red shaded region indicates a 2.5% probability, so the probability of lying outside the range (206.3, 293.5) is 5%.

Now, the 95% confidence interval for μ is

$$\bar{x} \pm t_{248,\,0.025}\sqrt{\frac{s^2}{249}}.$$

We have (from the `mean(netflix)` R output)

$$\bar{x} = \frac{1}{249}\sum_{i=1}^{249} x_i = 0.001426$$

and (from the `var(netflix)` R output)

$$s^2 = \frac{1}{248}\sum_{i=1}^{249}(x_i - \bar{x})^2 = 0.0008466.$$

so, substituting in the values for $\bar{x}, s^2$ and $t_{248,\,0.025}$, we compute the 95% confidence interval to be

$$(-0.002, 0.005).$$

For the population variance σ^2, the 95% confidence interval is

$$\left(\frac{248s^2}{\chi^2_{248,\,0.025}}, \frac{248s^2}{\chi^2_{248,\,0.975}}\right),$$

and substituting in the values we get (0.0007149, 0.0010172). We can take the square root to get a CI for the standard deviation: (0.027, 0.032).

To help understand how we might use these results see Figure 5.4, where compare the returns for the Netflix and GSK stocks. (Calculations for the GSK stocks confidence intervals are not given here, but are included in the tutorial booklet as an exercise.)

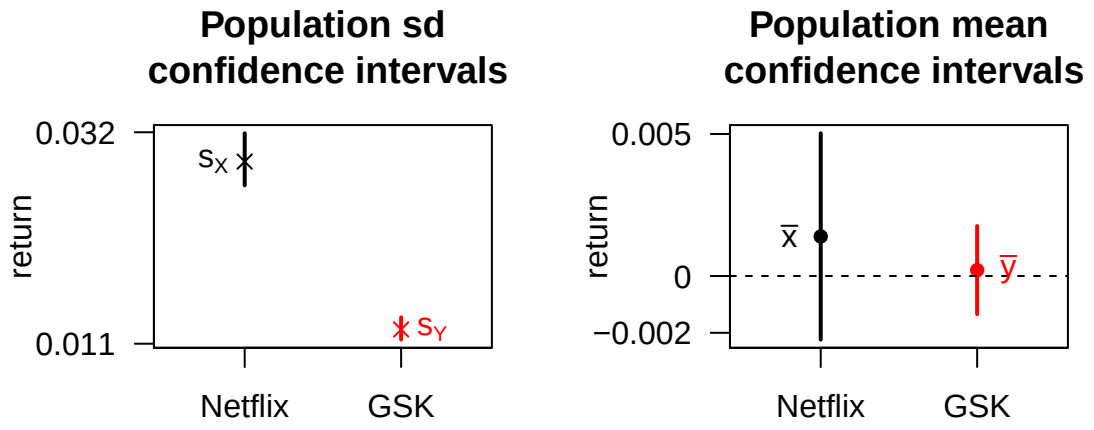


Figure 5.4: The left plot shows 95% confidence intervals for the standard deviations (volatilities). Here, we can be confident that the Netflix returns have a higher population standard deviation: investing in Netflix looks to be more risky. The right shows 95% confidence intervals for the mean returns for Netflix and GSK stocks. As a consequence of the higher standard deviation for Netflix, we are more uncertain about the population mean return compared with GSK, even though the sample sizes were the same. The Netflix population mean return could be much higher, but it could actually be lower than GSK's; this another way in which we can see the higher risk with Netflix.

We'll now justify these choices of interval estimates using the following result.

Theorem 5.2 (Property of a confidence interval)**.** Before *we get the data, a 95% confidence interval has a 95% chance of containing the true value of the parameter.*

Proof

We have

$$X_1, X_2, \ldots, X_n \overset{i.i.d}{\sim} N(\mu, \sigma^2),$$

and we consider the two random variables

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i, \tag{5.5}$$

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}^2). \tag{5.6}$$

(Recall that $\bar{X}$ and S^2 are random, because they are functions of the random variables $X_1, \ldots, X_n$). We define

$$T = \frac{\bar{X} - \mu}{\sqrt{S^2/n}}.$$

We first show that $T \sim t_{n-1}$, i.e the function T has the Student-t distribution with $n-1$ degrees of freedom. We write

$$T = \frac{\bar{X} - \mu}{\sqrt{\frac{\sigma^2}{n}}\sqrt{\frac{S^2}{\sigma^2} \times \frac{n-1}{n-1}}},$$

(where we can see that the σ^2 and $(n-1)$ terms cancel out, leaving us with the first expression for T above.) We can now write

$$T = \frac{Z}{\sqrt{Y/(n-1)}},$$

where

$$Z = \frac{\bar{X} - \mu}{\sqrt{\sigma^2/n}} \sim N(0,1),$$

which follows from equation (4.11), and

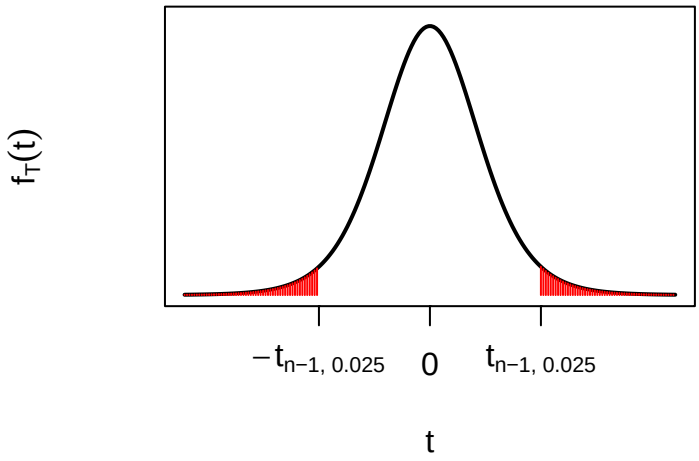
$$Y = \frac{S^2(n-1)}{\sigma^2} \sim \chi^2_{n-1},$$

which follows from equation (4.13). Then, applying Theorem 5.1, it follows that $T \sim t_{n-1}$.

Now, we have

$$P(-t_{n-1,\,0.025} \leq T \leq t_{n-1,\,0.025}) = 0.95,$$

which we visualise below.



Now we substitute in for T:

$$P\left(-t_{n-1,\,0.025} \leq \frac{\bar{X} - \mu}{\sqrt{\frac{S^2}{n}}} \leq t_{n-1,\,0.025}\right) \tag{5.7}$$

$$= 0.95. \tag{5.8}$$

Multiplying the inequality through by -1 we have

$$P\left(t_{n-1,\,0.025} \geq \frac{\mu - \bar{X}}{\sqrt{\frac{S^2}{n}}} \geq -t_{n-1,\,0.025}\right) = 0.95, \tag{5.9}$$

and then we can rearrange the inequalities (multiply by $\sqrt{\frac{S^2}{n}}$, then add $\bar{X}$) to get

$$P\left(\bar{X} + t_{n-1,\,0.025}\sqrt{\frac{S^2}{n}} \geq \mu \right. \tag{5.10}$$

$$\left. \geq \bar{X} - t_{n-1,\,0.025}\sqrt{\frac{S^2}{n}}\right) \tag{5.11}$$

$$= 0.95, \tag{5.12}$$

Hence **before** we get the data, there is a 95% chance that the interval

$$\left[\bar{X} - t_{n-1,\,0.025}\sqrt{\frac{S^2}{n}},\ \bar{X} + t_{n-1,\,0.025}\sqrt{\frac{S^2}{n}}\right]$$

will contain μ. This is the justification for using (5.1) as our interval estimate for μ: the probability that this approach will result in an interval that contains μ is high: 0.95.

We will illustrate this with the following simulation experiment. Using R, we will generate a sample of size 10 from a normal distribution with *known* parameters. We can then calculate the confidence interval for the mean, and see if it contains the true value or not. We will first do this once, using the $N(30, 25)$ distribution:

```
x <- rnorm(n = 10, mean = 30, sd = sqrt(25))
x
```

```
##  [1] 25.19 28.54 31.29 24.24 30.98 30.15 30.43 35.58 23.91 36.34
```

```
mean(x) - qt(0.975, 9) * sqrt(var(x) / 10)
```

```
## [1] 26.57
```

```
mean(x) + qt(0.975, 9) * sqrt(var(x) / 10)
```

```
## [1] 32.76
```

This gave a 95% confidence interval of (26.57, 32.76), which does contain the true value (30) in this instance. Now we'll repeat this 100 times, each time obtaining different random samples of size 10 from the $N(30, 25)$ distribution and each time calculating the confidence interval. The 100 confidence intervals are shown as horizontal lines in Figure 5.5.

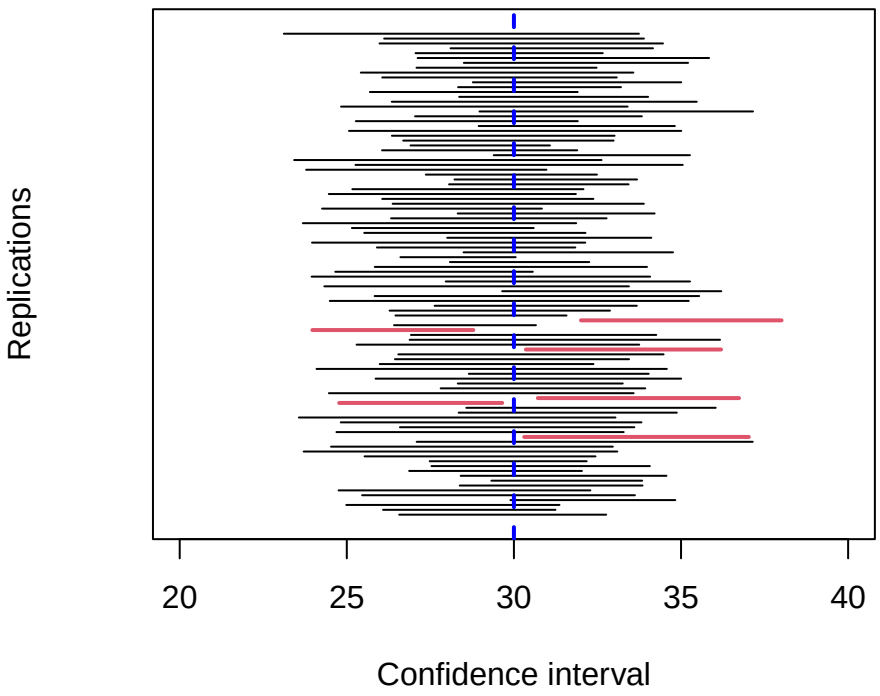


Figure 5.5: 95 % confidence intervals from one hundred separate samples of data. Before the data are obtained, we would expect 95 out of the 100 intervals to contain the true value of the mean. After the data are obtained, we see what 94 out of the 100 intervals did actually contain the true value.

The "95%" in "95% confidence interval" refers to a probability *before* getting the data.

5.3 Confidence interval for the probability parameter in a binomial distribution

Suppose we have

$$X \sim Bin(n, \theta)$$

Denoting the observed value of X by x, an *approximate* 95% confidence interval for θ is given by

$$p \pm z_{0.025}\sqrt{\frac{p(1-p)}{n}}, \tag{5.13}$$

where $p = x/n$ and $z_{0.025} = 1.96$ is the 97.5th percentile of the $N(0,1)$ distribution (so we are using the normal distribution rather than the t-distribution here.)

Example 5.2 (Confidence interval for a binomial probability parameter: Scottish independence opinion polls)**.** A survey has been conducted to estimate support for an independent Scotland[1]. 1067 voters in Scotland were asked: "Should Scotland be an independent country?". The responses were as follows: Yes: 43%, No: 45%, Don't know: 10%, Refused: 3%. Assuming each respondent was selected at random from the population of eligible voters, calculate an approximate 95% confidence interval

[1]See, for example, the opinion polls reported here

for the proportion of "Yes" voters in Scotland, ignoring the "Don't know" and "Refused" responses. What would the CI have been, assuming the same observed proportions, but with a sample size of 100 voters?

Solution.

The 95% confidence interval is

$$0.43 \pm 1.96\sqrt{\frac{0.43 \times 0.57}{1067}},$$

which gives (40%, 46%). Had the sample size been 100, with the proportions unchanged, the 95% CI would be

$$0.43 \pm 1.96\sqrt{\frac{0.43 \times 0.57}{100}},$$

which gives (33%, 53%). Arguably, this is too wide to be useful; in particular, the interest is going to be in whether the 'yes' vote exceeds 50%, and this interval spans 50%.

5.4 $100(1-\alpha)\%$ Confidence Intervals

We can consider other levels of confidence. In general, we use the expression "$100(1-\alpha)\%$ confidence interval", so, for example, choosing $\alpha = 0.01$ corresponds to a 99% confidence interval. We write the confidence intervals for the three cases we have considered as follows

- $100(1-\alpha)\%$ confidence interval for the mean of a normal distribution

$$\left[\bar{x} - \frac{s}{\sqrt{n}}t_{n-1;\alpha/2}, \quad \bar{x} + \frac{s}{\sqrt{n}}t_{n-1;\alpha/2}\right], \tag{5.14}$$

- $100(1-\alpha)\%$ confidence interval for the variance of a normal distribution

$$\left[\frac{(n-1)s^2}{\chi^2_{n-1;\alpha/2}}, \quad \frac{(n-1)s^2}{\chi^2_{n-1;1-\alpha/2}}\right], \tag{5.15}$$

- $100(1-\alpha)\%$ confidence interval for a binomial probability parameter

$$p \pm z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}}. \tag{5.16}$$

As we increase the confidence level (by decreasing α), the confidence intervals will become wider. The penalty for increasing the probability (before we get the data) that the interval will contain the true value is to report an interval that is less informative.

Example 5.3 (Confidence intervals: calculating a 99% confidence interval for a binomial probability parameter)**.** Calculate a 99% confidence interval for the population proportion of yes voters from the previous example. (43% yes voters from a sample of 1067), using the R output below. Only one of the three output values is relevant: you have to decide which.

```
qnorm(c(0.95, 0.99, 0.995), mean = 0, sd = 1)
```

```
## [1] 1.645 2.326 2.576
```

Solution

We want a 99% confidence interval, so in the general notation, we have a 100(1-α)% interval with $\alpha = 0.01$ The confidence interval is given by

$$0.43 \pm z_{0.01/2}\sqrt{\frac{0.43(1-0.43)}{1067}}. \tag{5.17}$$

We just need to know the value of $z_{0.01/2} = z_{0.005}$, which is the 99.5th percentile (not the 0.5th percentile!) of the standard normal distribution. From the R output above, this value is 2.576. Just to confirm this, we have

We display this in the plot below.

Substituting in 2.576 for $z_{0.005}$, we obtain the 99% confidence interval as (39%, 47%): slightly wider than the 95% interval, as it has to be. The price to pay for being more 'confident' is that we are less 'informative': we have to report a wider interval.

There is no point in attempting to produce a 100% confidence interval. For the mean μ of a normal distribution, for example, we have $t_{n-1;0} = \infty$, so the 100% confidence interval would be $(-\infty, \infty)$. That's clearly not helpful!

Chapter 6

Hypothesis testing: A-level recap

This chapter recaps the basic ideas in hypothesis testing that are covered at A-level (though we will *not* assume that you have studied hypothesis testing before). In the next few chapters we will cover some further hypothesis testing problems, as well as studying a general simulation approach, which will give you more insight into how hypothesis testing works.

In a hypothesis test, we make some assumption (the "hypothesis") about the distribution of the data, typically specifying the values of the parameters of the distribution, and then test whether the data 'contradict' this assumption.

There are two general approaches to hypothesis testing, which are very similar, but differ in how the conclusions are reported. We will refer to these as Neyman-Pearson testing, and Fisher's p-value method[1].

6.1 Hypothesis testing with the Neyman-Pearson approach

Suppose a company is developing a vegetarian substitute for minced beef, and is aiming for a product which is indistinguishable from meat. The substitute is to be tested in an experiment. Fifty volunteers will be each be given two portions of lasagna: one made with beef, and one with the vegetarian substitute, and asked to identify which lasagna is meat-free. The company will analyse the results of the experiment with a hypothesis test, and will make a decision about whether to continue with their product based on the results.

This is a scenario in which Neyman-Pearson testing could be used. The general procedure is as follows.

1. **Choose an appropriate statistical model and hypotheses**

Define X to be the number of people who correctly identify the meat-free lasagna. We might suppose that

$$X \sim Bin(50, \theta)$$

If the substitute is indistinguishable from minced beef, the volunteers would, in effect, be guessing with a probability of 0.5 of guessing correctly. Suitable null and alternative hypotheses would then be

$$H_0 : \theta = 0.5 \quad \text{(null hypothesis)}, \tag{6.1}$$

$$H_A : \theta \neq 0.5 \quad \text{(two-sided alternative hypothesis)}. \tag{6.2}$$

[1] devised by the statisticians Jerzy Neyman (1894-1981), Egon Pearson (1895-1980) and Sir Ronald Fisher (1890-1962).

2. **Choose the size or significance level of the test**

When stating the conclusion of the test, we will either state we "reject H_0" (conclude that H_0 is false), or we "do not reject H_0" (and then carry on as if H_0 is true). There are two ways, therefore, in which we could make the wrong conclusion. These are referred to as type I and type II errors.

Definition 6.1 (Type I error)**.** A Type I error is the mistake of *rejecting* the null hypothesis when the null hypothesis is actually *true*.

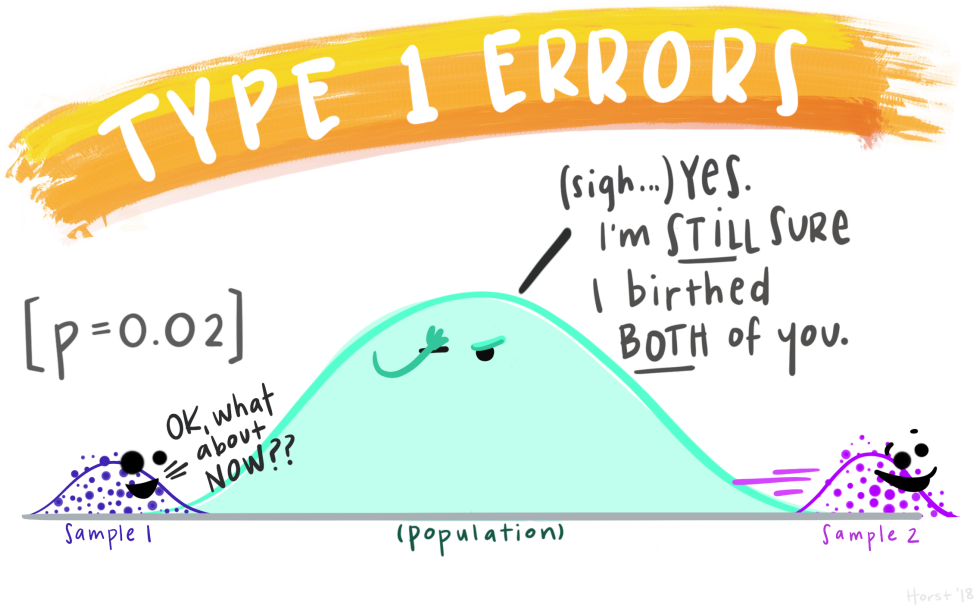


Figure 6.1: A type I error: falsely rejecting H_0. We think we have discovered something 'interesting' in our data, but have been deceived by random variation. Artwork by @allison_horst.

Definition 6.2 (Type II error)**.** A Type II error is the mistake of *failing to reject* the null hypothesis when the null hypothesis is actually *false*.

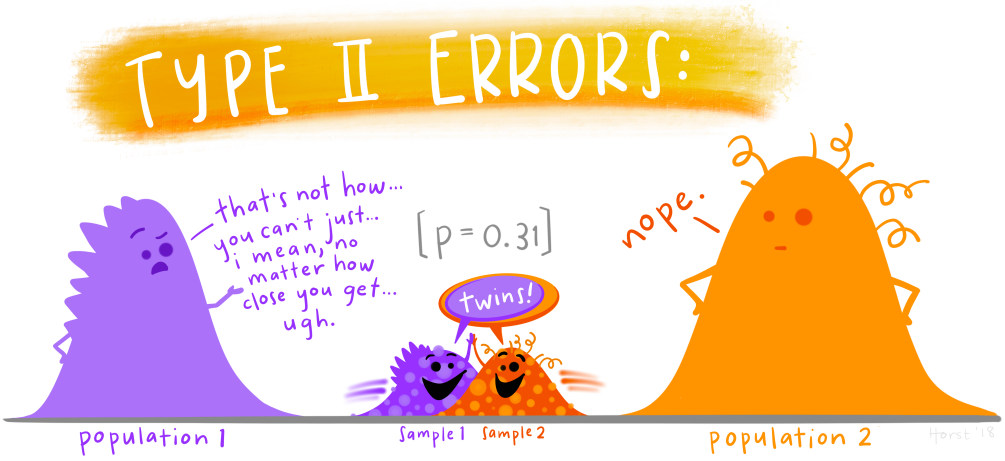


Figure 6.2: A type II error: failing to reject H_0 when H_0 is really false. Here, random variation has hidden a potentially interesting discovery. This can result from having too small a sample size. Artwork by @allison_horst.

Definition 6.3 (Size / level of significance)**.** The size of a test is the probability, before we get our data, that we would make a Type I error. The size of a test is also known as the level of significance. The size/level of significance is often denoted by α.

In Neyman-Pearson testing we choose, in advance, the size of test. A common choice of size/significance level is 5%, so the probability of a Type I error would be 0.05. Why not choose 0%, so that a Type I error is impossible? The only way to make Type I errors *impossible* is to refuse ever to reject the null hypothesis, but this then *increases* the risk of a Type II error; we have to trade off risks of the two error types. Choosing a small value such as 5% is a compromise.

3. **Choose a test statistic**

> A test statistic measures 'how different' the data are from what we would expect under H_0.

In our example, we use the test statistic

$$Z = \frac{\frac{X}{n} - \theta_0}{\sqrt{\theta_0(1-\theta_0)/n}}, \tag{6.3}$$

with n the sample size, and θ_0 the hypothesised values of θ under H_0 (so in the example, we have $n = 50$ and $\theta_0 = 0.5$). This measures how far the observed proportion of correct responses is from θ_0 (the proportion we'd expect if H_0 were true), scaled by the standard deviation of X/n (again, if H_0 were true).

> We'd expect (the absolute value of) a test statistic to be small if H_0 is true, and to be relatively large if H_A is true.

4. **Identify the critical region**

We now find a critical region C such that

- a value of Z in the critical region would correspond to a large difference between X/n and θ_0;

- the probability of Z falling in the critical region, if H_0 were true, would be 0.05 (the size/significance level).

Using the normal approximation to the binomial distribution, we suppose that $Z \sim N(0,1)$ and so the critical region is $(-\infty, -1.96] \cup [1.96, \infty)$

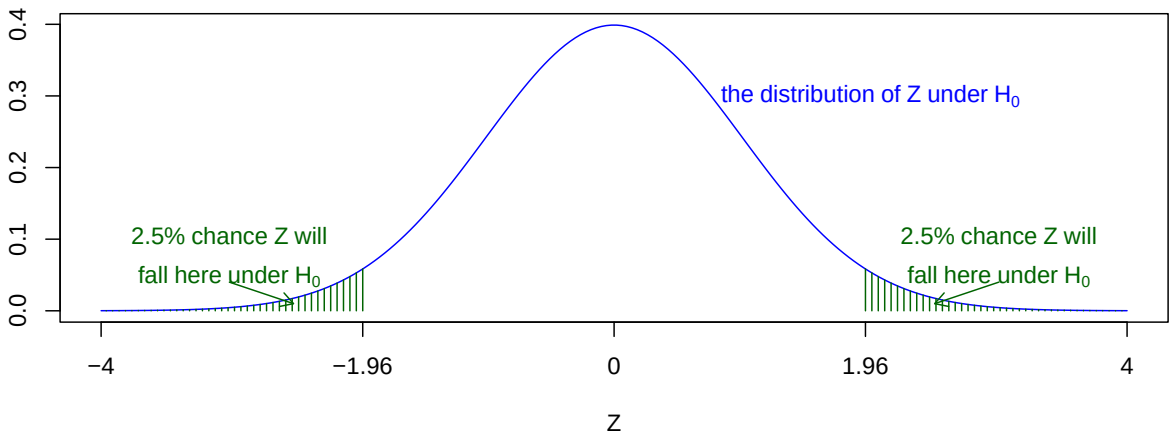


Figure 6.3: The distribution of the test statistic assuming H_0 is true, and the 5% critical region shaded in green. There is only a 5% chance of the test statistic falling in this region if H_0 is true: we will reject H_0 if the test statistic falls in this region.

5. **Compute the test statistic for the observed data, and state the conclusion**

If the observed value of Z falls in the critical region, we declare that we reject H_0 (at the 5% level of significance); otherwise, we declare that we do not reject H_0. For example, if 40 out of 50 people correctly identified the meat-free lasagna, the observed test statistic, denoted by z_{obs} would be

$$z_{obs} = \frac{\frac{40}{50} - 0.5}{\sqrt{0.5 \times (1-0.5)/50}} = 4.24,$$

which does lie in the critical region, so H_0 would be rejected. The company may then decide that their substitute hasn't achieved the taste they want, and they may try something else.

6.1.1 One-sided and two-sided alternative hypotheses

We used an alternative hypothesis

$$H_A : \theta \neq 0.5.$$

This is **two-sided** because we would want to reject $H_0 : \theta = 0.5$ if either $\theta > 0.5$ or $\theta < 0.5$. **One-sided** alternative hypotheses would be

$$H_A : \theta > 0.5,$$

or

$$H_A : \theta < 0.5.$$

In some situations, it may appear that a one-sided alternative hypothesis is more suitable, e.g. H_0 : "the drug has no effect on blood pressure, on average" and H_A : "the drug lowers blood pressure, on average" (if the aim of the drug was to lower blood pressure). However, there is an argument in this situation for **always using a two-sided alternative**.

- If the drug had the opposite effect to that desired, we would still want to know.
- Using a one-sided alternative makes the critical region larger in the area of interest; H_0 can be rejected with a smaller observed effect of the drug.

6.2 Fisher's p-value method

In the Neyman-Pearson approach to hypothesis testing, the conclusion is stated in terms of "reject H_0", or "do not reject H_0". We would use this where there is a clear **decision** to be made after the test. Sometimes, however, we are just interested in whether data supports a particular hypothesis or not; there is no decision or action that follows the test.

> A hypothesis test can not **prove** whether a hypothesis or not. Rather than declaring whether a hypothesis been "rejected", it might be preferable instead to report the strength of evidence provided by an experiment.

Consider again the example of the vegetarian minced-beef substitute. Suppose the product is already on the market, and a consumer TV show does the same experiment to see if people if can taste the difference. There is no 'decision' to be made afterwards; the experiment is done out of public interest.

We have the same model and hypotheses as before. Defining X to be the number of people (out of 50) correctly identifying the vegetarian substitute, we suppose $X \sim Bin(50, \theta)$, with

$$H_0 : \theta = 0.5, \tag{6.4}$$
$$H_A : \theta \neq 0.5, \tag{6.5}$$

so that under H_0, people are just guessing. Now consider three scenarios:

Scenario	Data	Test Statistic
A	32 people out of 50 guess correctly	2.06
B	31 people out of 50 guess correctly	1.75
C	40 people out of 50 guess correctly	5.30

If we were using Neyman-Pearson with a test of size 0.05, the critical region for the test statistic would be $(-\infty, -1.96] \cup [1.96, \infty)$; the test statistics would lie in the 5% critical region in scenarios A and C, but not in scenario B.

- Comparing scenarios A and B, the results were the nearly the same: only one more person correctly identified the meat substitute in scenario A. Should we really be drawing different conclusions in these two scenarios?
- Comparing scenarios A and C, the evidence seems more persuasive in scenario C. Shouldn't we report this somehow?

In Fisher's p-value method, instead of declaring whether H_0 is reject or not, we describe **the strength of the evidence against** H_0, by reporting the **p-value**.

The p-value is a probability and, informally, describes how 'surprising' the observed data are, assuming H_0 to be true.

- If the p-value is small, it means the data are not what we would expect to see under H_0. The smaller the p-value, the stronger the evidence *against* H_0.
- If the p-value is large, the data are consistent with what we'd expect under H_0 (but this is not the same as saying we have evidence *in favour* of H_0 being true).

Definition 6.4 (p-value)**.** For a test statistic T, with observed value t_{obs} and a two-sided alternative hypothesis, we define the p-value as

$$P(|T| \geq |t_{obs}|), \tag{6.6}$$

calculated for the distribution of T under H_0.

Continuing the example, our test statistic is denoted by Z, assumed to have a $N(0,1)$ distribution if H_0 is true. The p-values in the three scenarios would be:

$$\begin{aligned}
\text{Scenario A:} \quad & P(Z \leq -2.06) + P(Z \geq 2.06) \simeq 0.04, \\
\text{Scenario B:} \quad & P(Z \leq -1.75) + P(Z \geq 1.75) \simeq 0.08, \\
\text{Scenario C:} \quad & P(Z \leq -5.30) + P(Z \geq 5.30) \simeq 1.2 \times 10^{-7}.
\end{aligned}$$

Note that the p-value is much smaller in Scenario C than in A: the evidence against H_0 is stronger.

We obtain the numerical probabilities using R, noting that for $Z \sim N(0,1)$, we have $P(Z \leq -x) + P(Z \geq x) = 2P(Z \leq -x)$. For example

```
2 * pnorm(-2.06)
```

```
## [1] 0.0394
```

The p-value is much smaller in Scenario C, compared with A, and so we can say that the evidence against H_0 is much stronger in Scenario C. We visualise the p-value in Scenario A below.

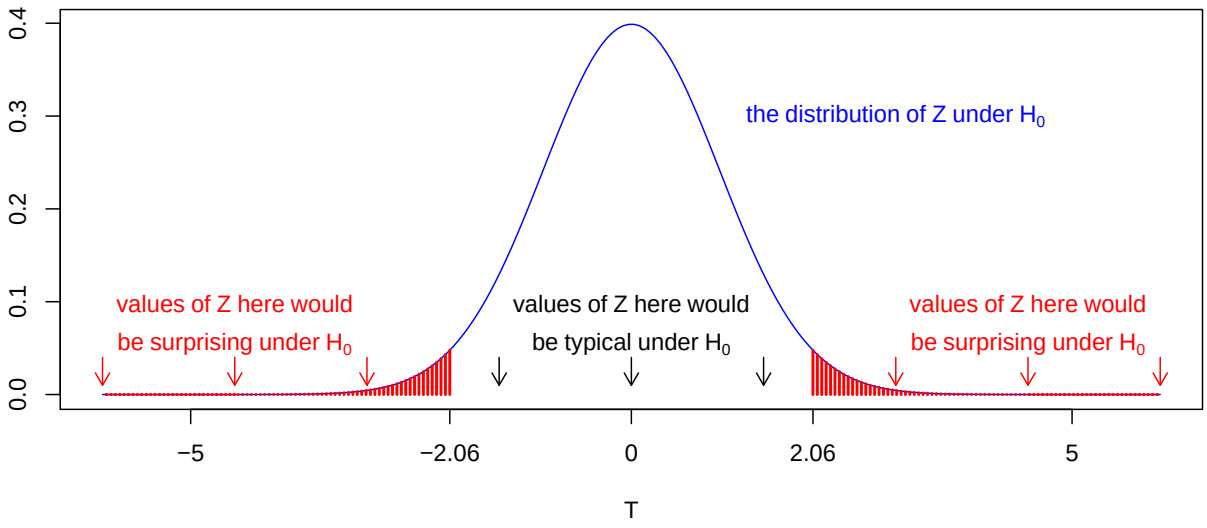


Figure 6.4: The observed test statistic was 2.06. Is this a suprising value of Z under H_0? We report how surprising this value is in terms of how likely we are to get an (absolute) value as large as 2.06, assuming H_0 to be true. This is the p-value, and is represented by the red shaded area.

6.2.1 What counts as a small p-value?

This varies between scientific fields. In medical research, a p-value of 0.05 or smaller would typically count as 'significant' evidence against the null hypothesis. If a scientist wants to claim a new discovery, and publish the results of his/her experiment in an academic journal, some journals will require a p-value less than 0.05 for the article to be published, although one journal banned this practice[2]. Particle physicists are rather more demanding! They require a p-value smaller than 0.003 for "evidence of a particle", and smaller than 0.0000003 for a "discovery"[3].

For this module we will use the following convention

p-value	Interpretation
$p > 0.05$	No evidence against H_0
$0.05 \geq p > 0.01$	Weak evidence against H_0
$0.01 \geq p > 0.001$	Strong evidence against H_0
$0.001 \geq p$	Very strong evidence against H_0

In particular, for p-values just below 0.05, a recommendation would be to repeat the experiment to look for confirmation.

6.3 Relationship between the Neyman-Pearson and p-value methods

Note that if the p-value is less than 0.05, we can deduce that the test statistic must lie in the 5% critical region. Hence some people use a combination of both methods, and say things like, "The p-value is

[2]https://www.statslife.org.uk/news/2116-academic-journal-bans-p-value-significance-test
[3]https://blogs.scientificamerican.com/observations/five-sigmawhats-that/

less than 0.05, so we have statistically significant evidence against H_0 at the 5% level." Reporting the p-value gives a little more information: we are saying how strong the evidence is against H_0, and not simply whether H_0 is rejected or not. We illustrate this in Figure 6.5.

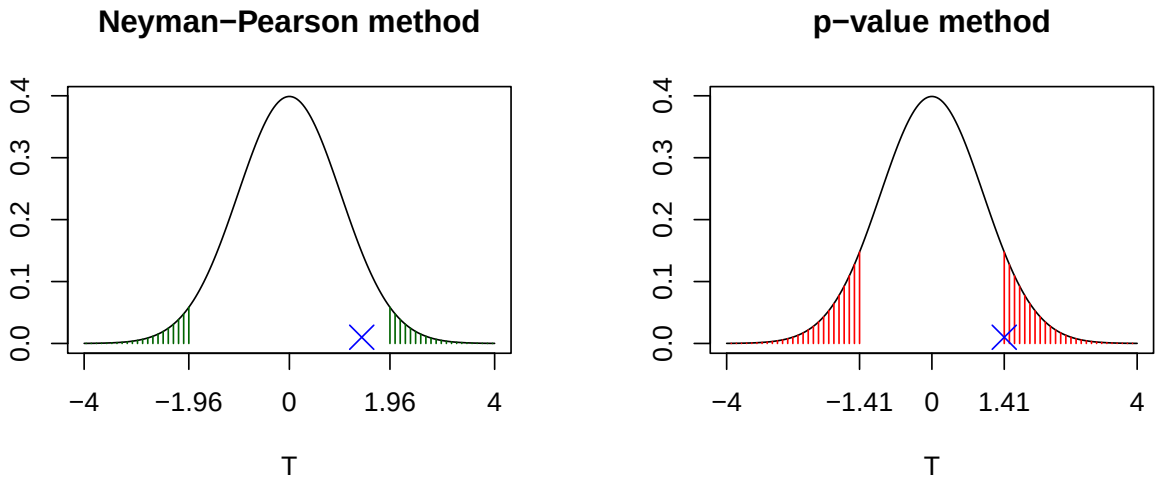


Figure 6.5: Suppose our observed test statistic was 1.41, as shown by the blue cross. For the Neyman-Pearson method, with a test of size 0.05, we determine the critical region which has a 5% chance of containing the test statistic T, assuming H_0 is true. This is shown as the green shaded area; this area is 0.05. For the p-value method, we calculate the probability that T would be as or more extreme as our observed test statistic t_{obs}, assuming H_0 is true. This is shown as the red shaded area. If the p-value (red shaded area) is greater than 0.05, we can deduce that the test statistic t_{obs} cannot lie in the 5% critical region.

6.4 One-sample t-test

One of the most common tests undertaken in practice is the one-sample t-test. Here, $X \sim N(\mu, \sigma^2)$, where the value of the population variance σ^2 is unknown, and we test the value of the population mean μ. Our testing procedure in this case is as follows:

1. State your null and alternative hypotheses:

$$H_0 : \ \mu = \mu_0 \qquad\qquad H_A : \ \mu \neq \mu_0$$

2. Choose a test statistic T:

$$T = \frac{\bar{X} - \mu_0}{\sqrt{S^2/n}}$$

3. Identify the distribution of T assuming H_0 is true:

$$T \sim t_{n-1}$$

4. Compute the value of the test statistic for the observed data, t_{obs}.

5. Conclude the test: "reject" or "do not reject" H_0, **or**, report the p-value (strength of evidence against H_0).

> If the population variance σ^2 **is known**, then this test becomes a 'one-sample Z-test', where we denote the test statistic as Z and use σ^2 directly in the test statistic formula (rather than estimating it with S). In this case, under H_0, $Z \sim N(0,1)$.

Example 6.1 (One sample t-test: Energy drink company claim)**.** A drinks company claims that the energy drinks they produce contain 20cl on average per can. A food standards authority decides to test this claim. They sample 30 cans (bought at different retailers over a one month period) and measure the quantity of energy drink (X) inside each can. A summary of the observed data is as follows:

$$\bar{x} = 20.12, \qquad s^2 = 0.179.$$

Perform a one-sample t-test to test the claim of the drinks company.

Useful R output:

```
qt(c(0.95, 0.975, 0.99, 0.995), df = 29)
```

```
## [1] 1.699 2.045 2.462 2.756
```

Solution

Let the random variable X be the quantity of energy drink inside a can produced by the company, and assume that $X \sim N(\mu_X, \sigma_X^2)$, where μ_X represents the population mean quantity of X and σ_X^2 represents the population variance of X, which is unknown. The company's claim is that $\mu_X = 20$cl. We test the claim as follows:

Our hypotheses are:

$$H_0 : \ \mu_X = 20; \qquad \text{v's} \qquad H_A : \ \mu_X \neq 20$$

Our test statistic T is:

$$T = \frac{\bar{X} - \mu_0}{\sqrt{S^2/n}},$$

where $\mu_0 = 20$.

The Food Standards Authority has collected a sample of size $n = 30$. Hence, assuming that our null hypothesis is true, the distribution of the test statistic T is:

$$T \sim t_{29}.$$

The sample statistics are: $n = 30$, $\bar{x} = 20.12$, and $s^2 = 0.179$. Using these values, our observed test statistic, t_{obs}, is:

$$t_{obs} = \frac{20.12 - 20}{\sqrt{0.179^2/30}} = \frac{0.12}{0.0772442} = 1.55 \text{ (2dp)}.$$

Conclude the test:

1. Neyman-Pearson approach: For a 2-sided test at the 5% significance level ($\alpha = 0.05$), the critical value is the t-quantile: $t_{29,0.025} = 2.045230$ (from the R output). Hence our critical region is given by:
$$(-\infty\ ,\ -2.045230] \cup [2.045230\ ,\ \infty)\,.$$
The observed test statistic of 1.55 is **not** inside this critical region, so we do not reject H_0 and conclude in favour of H_0 at the 5% significance level. We conclude that the cans do contain 20cl of drink on average, and that no action should be taken against the drinks company by the Food Standards Authority.

2. p-value approach: Under the p-value approach, we assess the value of the p-value in this case using the R output. We have:
$$p\text{-value} = P\left(|T| \geq |t_{obs}|\right) > 0.1,$$
as $t_{obs} = 1.55 < 2.045230$. (Using R, the actual p-value is 0.1291.) Hence, we have no evidence against the null hypothesis H_0, and we conclude in favour of H_0, as above.

6.5 Which hypothesis test do I use for...?

There are a large number of hypothesis tests covering a range of situations. Over the next few chapters we will consider three (studying more than this would be tedious!):

1. comparing two means;
2. comparing two proportions;
3. analysing contingency table data.

You will see that the general approach is the same in each case. All that change are

- the test statistic that is computed;
- the distribution of the test statistic under the null hypothesis.

Once you have understood how things work in general, you should be confident in tackling any hypothesis testing problem: search for the problem online (or in a textbook), identify the choice of test statistic and its distribution under H_0, and then you should find the implementation straightforward.

Chapter 7

Hypothesis testing: comparing two population means

In this chapter we will use hypothesis testing to compare two populations and see if they have different population means. We will use two different methods: a computer simulation method, and an analytical method known as the **two-sample t-test**.

7.1 Example: can imagining eating food make you eat less?

Morewedge et al. (2010)[1] conducted an experiment to test whether imagining eating food can make one eat less, when offered the same food item to eat. The experiment was repeated by Camerer et al. (2018)[2], and we use their data here.

- 96 student volunteers were recruited, and split into two groups:
    1. In the **control group**, the participants were each shown a picture of a bowl filled with thirty-three 20-cent coins. They were asked to imagine inserting the coins, one at a time, into a parking meter.
    2. In the **treatment group**, the participants were each shown a picture of a bowl filled with three 20-cent coins. They were asked to imagine inserting the coins, one at a time, into a parking meter. They were then shown a picture a bowl containing 30 M&Ms, and they were asked to imagine eating the M&Ms, one at a time.
- All the participants were then given an actual bowl of M&Ms to eat (containing 40g in total). They were told they were doing a taste test, and were told to eat as much or a little they liked.
- Each participant did the experiment in a private cubicle, so no-one watched them eat, but the amount eaten was recorded once they had finished.

Histograms of the amounts eaten for the two groups are shown in Figure 7.1

[1] Morewedge, C. K., Huh, Y. E. & Vosgerau, J. Thought for food: imagined consumption reduces actual consumption. Science 330, 1530–1533 (2010).

[2] Camerer, C. F., Dreber, A., Holzmeister, F., Ho, T.-H., Huber, J., Johannesson, M.,… Wu, H. (2018). Evaluating the replicability of social science experiments in Nature and Science between 2010 and 2015. Nature Human Behaviour, 2(9), 637–644. https://doi.org/10.1038/s41562-018-0399-z

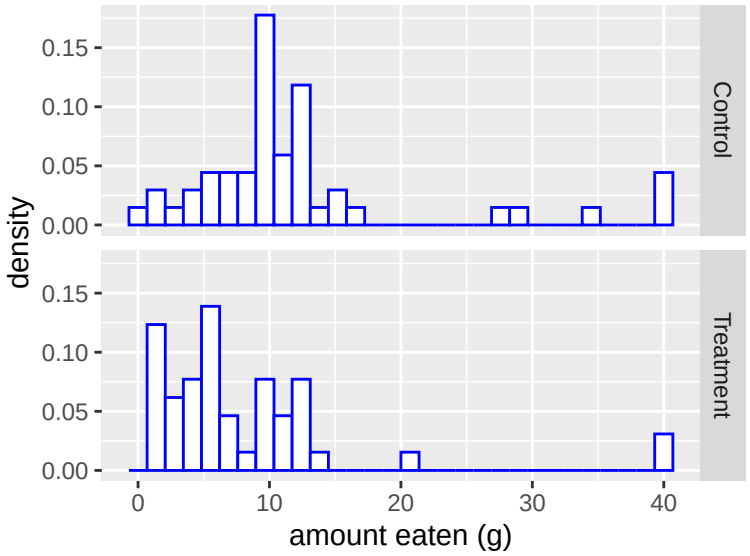


Figure 7.1: Histograms showing the amount of chocolate eaten in each group. The treatment group had to first imagine eating chocolate, before they were given anything to eat. Did it make them want to eat less?

7.1.1 The hypotheses

We define μ_X to be the population mean quantity eaten that would be eaten under the control conditions (*not* imagining eating M&Ms), and μ_Y to be the population mean quantity eaten that would be eaten under the treatment conditions (imagining eating M&Ms). The null hypothesis is that imagining eating M&Ms has no effect on consumption:

$$H_0 : \mu_X = \mu_Y$$

We choose a two-sided alternative

$$H_A : \mu_X \neq \mu_Y$$

as we would be interested if imagining eating M&Ms either resulted in the participants eating more, or eating less on average.

We define:

- $x_1, \ldots, x_n$: the n control group observations, with sample mean $\bar{x}$ and sample variance s_X^2;
- $y_1, \ldots, y_m$: the m treatment group observations, with sample mean $\bar{y}$ and sample variance s_Y^2.

In this example, we have $n = 49$ and $m = 47$.

7.1.2 A test statistic

We measure the difference in mean consumption with the test statistic

$$t_{obs} = \frac{\bar{x} - \bar{y}}{\sqrt{\frac{s_X^2}{n} + \frac{s_Y^2}{m}}}$$

so that we scale the difference in means by how much variation there is in the amounts the individuals ate.

In R, the observations are stored in the vectors `control` and `treatment`, and we compute t_{obs} as follows

```
(mean(control) - mean (treatment))/
  sqrt(var(control)/49 + var(treatment)/47)
```

```
## [1] 2.398
```

(We will round this to 2.4 from now on.)

7.2 Hypothesis testing using simulation

All hypothesis testing problems involve understanding what sort of data could arise *purely by chance*, where "purely by chance" is described by the null hypothesis. In the current example, could the difference in mean consumption have arisen purely by chance?

We will first use a computer simulation technique to investigate this. Let's look at the first few observations in the treatment group

```
treatment[1:4]
```

```
## [1] 12  7  5  8
```

We see that person 1 ate 12g, person 2 ate 7g and so on. Suppose the null hypothesis is true, and treatment has no effect. We might then suppose that, had person 1 been allocated to the control group instead, *it would have made no difference* to how much person 1 ate: he/she would still have eaten 12g. So maybe the difference is purely because more 'hungry' volunteers were randomly allocated to the control group than the treatment group.

Could this have happened by chance? We can investigate this as follows, bearing in mind that the randomness we are investigating here is the random allocation of volunteers to groups, and how that could produce unequal mean consumption.

1. We first combine all the treatment and control observations into a single vector called `everyone`.

```
everyone <- c(treatment, control)
```

so to see all the observations:

```
everyone
```

```
##  [1] 12  7  5  8 40  4  4 11  7  5 13 11 10  4  7  2 13 12  2  2  9 10  3  2  1
## [26]  2  3  4  9  3  3 14  4  9  6  5  2 40  6 11 21  5  2  5  5  6 12  2  8 11
## [51]  0 40 12 40 11 15  2 12  4  9 11 11 13  8 12  9 10 12 10  4  7 10  8  9 14
## [76]  6  7  5  9 17 13  9 10 10 13 34  7 12  6 10 28 15  3 29 40 10
```

2. We now randomly allocate each person into either the treatment group or the control group.

We first jumble up the order of the observations in `everyone`, using the `sample()` command

```
everyoneJumbled <- sample(everyone)
```

This is what we got:

```
everyoneJumbled
```

```
##  [1] 11  5  3 12 10 21 11  2 14  5 13  7 12  1  8  9  3 10 11  3  5  9  8 12  7
## [26] 11 12 10  8  2 10  6 40  4 40  4  2  6  9  2  3  9  9 28 14 40 12  9 10  2
## [51]  2 10 12  6 40  9  7 11 10  2  7  4  7 11 12  2 13  5  6 34  5  2  3 13 15
## [76] 10  4 15  5 40 13 11 12  9  7  4 10 29 13  4  8  5  6 17  4  0
```

and every time we use the `sample()` command, the observations in `everyone` will be jumbled up in a different order.

3. We now extract the first 47 elements to be a new set of random treatment observations:

```
newTreatment <- everyoneJumbled[1:47]
```

and the remaining 49 elements to be a new set of random control observations:

```
newControl <- everyoneJumbled[48:96]
```

We've reallocated each person into either the treatment group or the control group, and assuming H_0 is true, that the treatment has no effect, 'switching' someone from one group to the other *wouldn't change how much that person would eat.*

4. Now we'll see how different the mean consumptions would have been (scaled by the standard deviations)

```
(mean(newTreatment) - mean(newControl))/
    sqrt(var(newTreatment)/47 + var(newControl)/49)
```

```
## [1] 0.2098
```

This gives a much smaller test statistic.

Now we'll repeat the process lots of times, and see how easy it is to generate a test statistic as large as the one we got ($t_{obs} = 2.4$):

```
testStatistics <- rep(0, 100000)
everyone <- c(treatment, control)

for(i in 1:100000){
  everyoneJumbled <- sample(everyone)
  newTreatment <- everyoneJumbled[1:47]
  newControl <- everyoneJumbled[48:96]
  testStatistics[i] <- (mean(newTreatment) - mean(newControl))/
    sqrt(var(newTreatment)/47 + var(newControl)/49)
}
```

We now count how many times we got a test statistic larger (in absolute) value than 2.4:

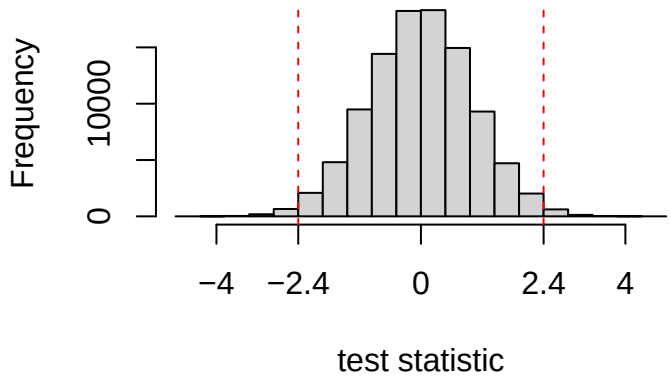
```
sum(abs(testStatistics) >= 2.4)
```

```
## [1] 1650
```

We see that about only 1650 times out of 100,000 did we generate test statistics (scaled differences between the sample means) as large as 2.4: the probability of producing a difference between the groups as large as this *purely by random chance* is about 2%.

To visualise this, we'll plot a histogram of our random test statistics:

We'll draw a histogram of the randomly generated test statistics:



We can see a symmetrical distribution around 0, with most smaller in absolute value than 2.4

In effect, we've computed a p-value: we've used simulation to estimate the probability, assuming H_0 to be true, of getting a test statistic as large as the one we observed.

7.3 The two-sample t test

As an alternative to the computer simulation, we can use an analytical approach, known as the two-sample t-test.

Before the experiment has been conducted define X_i to be the amount that the i-th participant in the control group will eat, and Y_i to be the amount i-th participant in the treatment group will eat. Before the experiment has been conducted, we can think of X_i and Y_i as random variables: their values are not yet known.

1. **The model and hypotheses**

We now suppose that

$$X_1, \ldots, X_n \overset{i.i.d}{\sim} N(\mu_X, \sigma_X^2),$$
$$Y_1, \ldots, Y_m \overset{i.i.d}{\sim} N(\mu_Y, \sigma_Y^2),$$

so that the population mean amounts eaten under the treatment and control conditions would be μ_X and μ_Y. We consider the hypotheses

$$H_0 : \mu_X = \mu_Y,$$
$$H_A : \mu_X \neq \mu_Y,$$

so the null hypothesis is that there is no difference between the mean amount eaten under either condition (it doesn't matter what the participants imagine doing before they eat.)

2. **The test statistic, and its distribution under H_0**

We use the test statistic

$$T = \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{S_X^2}{n} + \frac{S_Y^2}{m}}} \tag{7.1}$$

Assuming H_0 is true, then approximately

$$T \sim t_\nu,$$

i.e T has a student t distribution with ν degrees of freedom.

> As long as the sample sizes are moderately large (say at least 30 per group), this approximation is usually safe to use, even if the individual observations are *not* normally distributed (the Central Limit Theorem comes into play here).

We will determine ν from the data: we use what is known as the **Welch approximation**

$$\nu = \frac{\left(\frac{s_X^2}{n} + \frac{s_Y^2}{m}\right)^2}{\frac{(s_X^2/n)^2}{n-1} + \frac{(s_Y^2/m)^2}{m-1}}. \tag{7.2}$$

3. **Computing the p-value**

We compute the value of the test statistic for the observed data

$$t_{obs} = \frac{\bar{x} - \bar{y}}{\sqrt{\frac{s_x^2}{49} + \frac{s_y^2}{47}}} = 2.389.$$

and, using the Welch approximation (7.2), we compute the degrees of freedom to be

$$\nu = \frac{\left(\frac{s_X^2}{49} + \frac{s_Y^2}{47}\right)^2}{\frac{(s_X^2/49)^2}{49-1} + \frac{(s_Y^2/47)^2}{47-1}} \simeq 93.$$

To calculate the p-value, we calculate

$$\begin{aligned} &P(|T| \geq |t_{obs}| \,|H_0 \text{ true}) \\ &= P(|T| \geq 2.4 | H_0 \text{ true}) \\ &= P(T \leq -2.4 | H_0 \text{ true}) \\ &+ P(T \geq 2.4 | H_0 \text{ true}), \end{aligned}$$

where T has the t_{93} distribution under H_0. The p-value is shown as the red shaded area below. We also show the histogram of test statistics obtained using the simulation method.

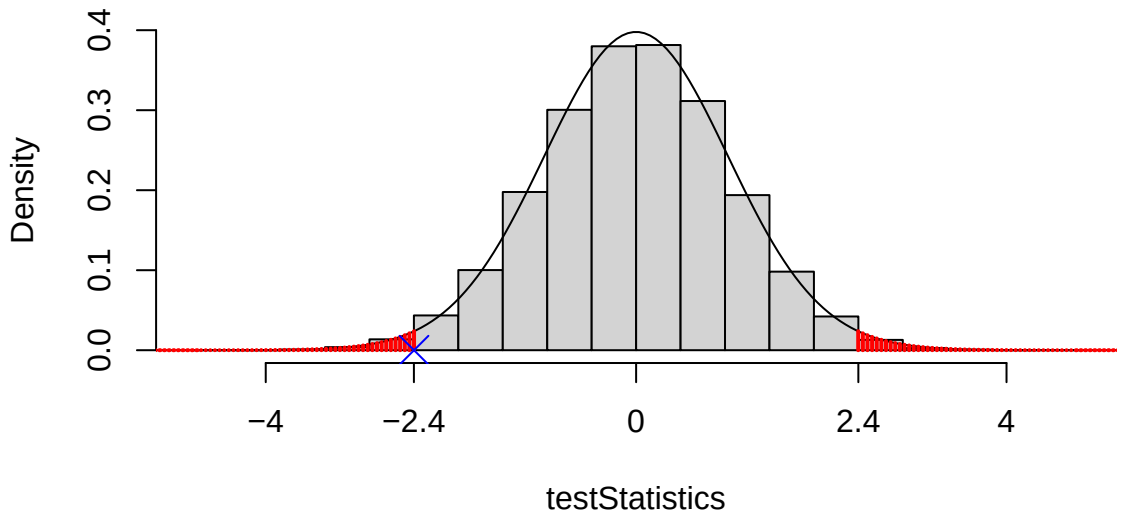


Figure 7.2: The observed test statistic (shown as the blue cross) was -2.4. For the p-value, we want the probability that T would be as extreme as this: either less than -2.4, or greater than 2.4. This probability is shown as the red shaded area. For comparison, we also show the histogram of test statistics from the simulation method. Note the close agreement with the t-distribution.

To calculate the p-value using R: we want

$$P(T \leq -2.4) + P(T \geq 2.4) \tag{7.3}$$
$$= 2 \times P(T \leq -2.4) \tag{7.4}$$

so the R command to get the p-value is

```
2 * pt(-2.4, 93)
```

```
## [1] 0.01839
```

hence the p-value approximately 0.02.

To interpret this, we can say that the experiment has provide some evidence that imagining eating food can reduce how much you want to eat! The evidence is not very strong, but this experiment was a *replication* of an earlier study: two independent studies found the same effect, so taken together, the evidence is more convincing.

7.3.1 The two-sample t-test with the Neyman-Pearson method

If using the Neyman-Pearson method, we would instead (after choosing the size of the test) identify the critical region. For a test of size 0.05, and a two-sided alternative hypothesis, we would need the 2.5th and 97.5th percentiles of the t_ν distribution. Continuing the example, in R, we would do

```
qt(0.975, 93)
```

```
## [1] 1.986
```

so the critical region would be $(-\infty, -1.99] \cup [1.99, \infty)$, as shown below.

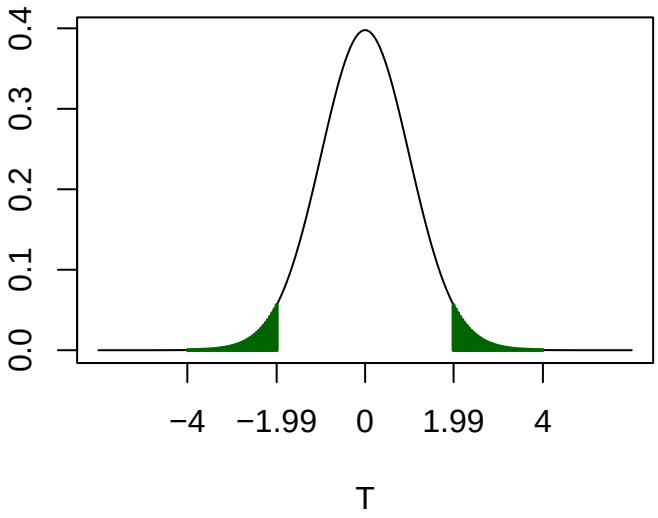
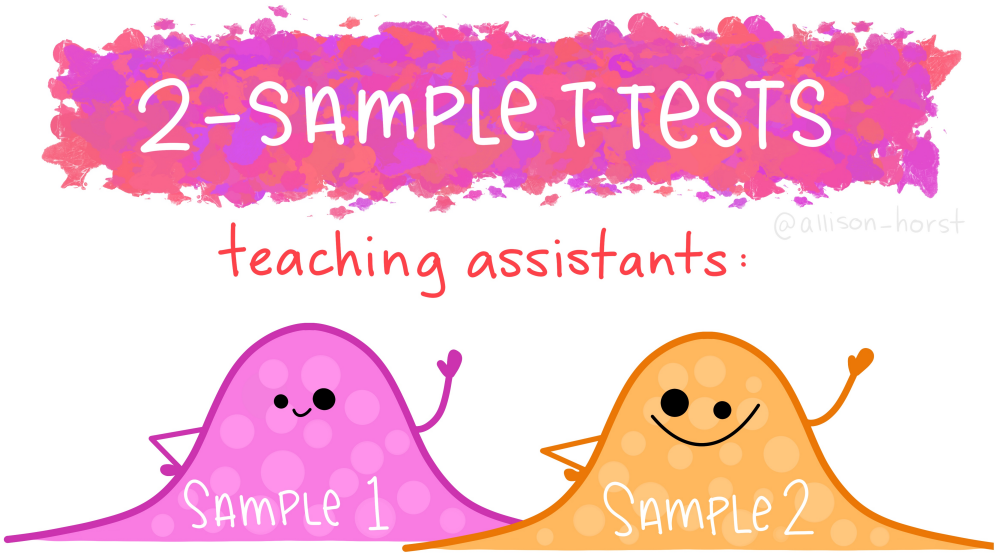


Figure 7.3: The critical region for a test of size 0.05 is shown as by the green shaded area. (The observed test statistic was 2.4, which does fall in this region, so if using the Neyman-Pearson framework, we would conclude that H_0 is rejected at the 5% level of significance.

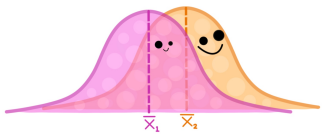
7.3.2 An illustrated guide

With thanks again to @allison_horst, below is an illustrated guide to two-sample t tests.

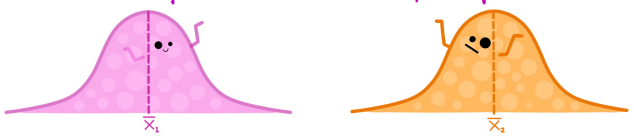


LET'S START HERE: if random samples are drawn from populations w/ the same mean... (i.e. the same population)

Then it is more likely that the 2 sample means will be close together...



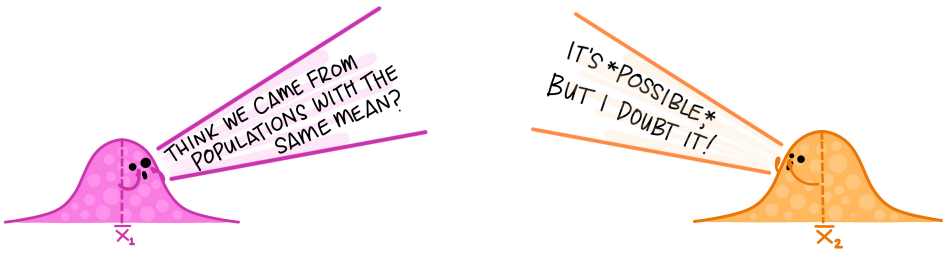
...and it is less likely (but always possible!) that the sample means will be far apart.



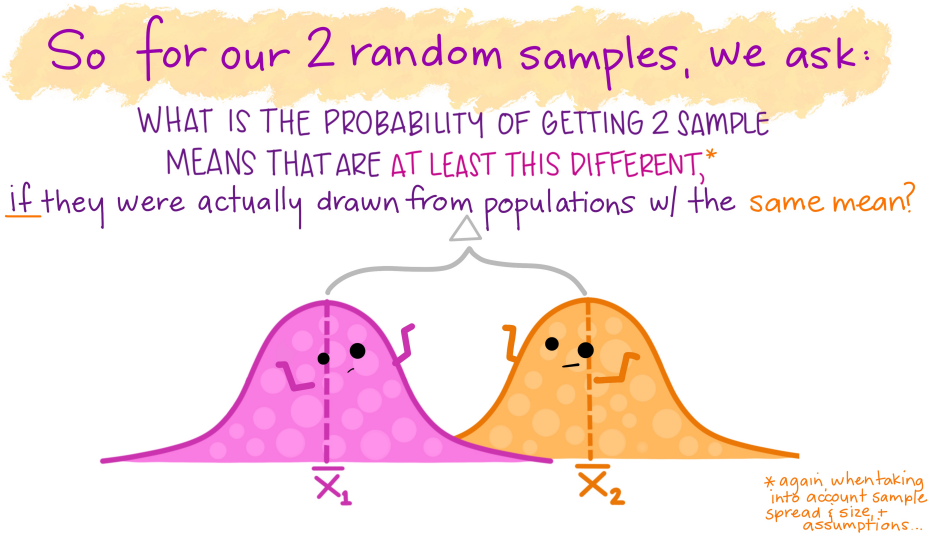
@allison_horst

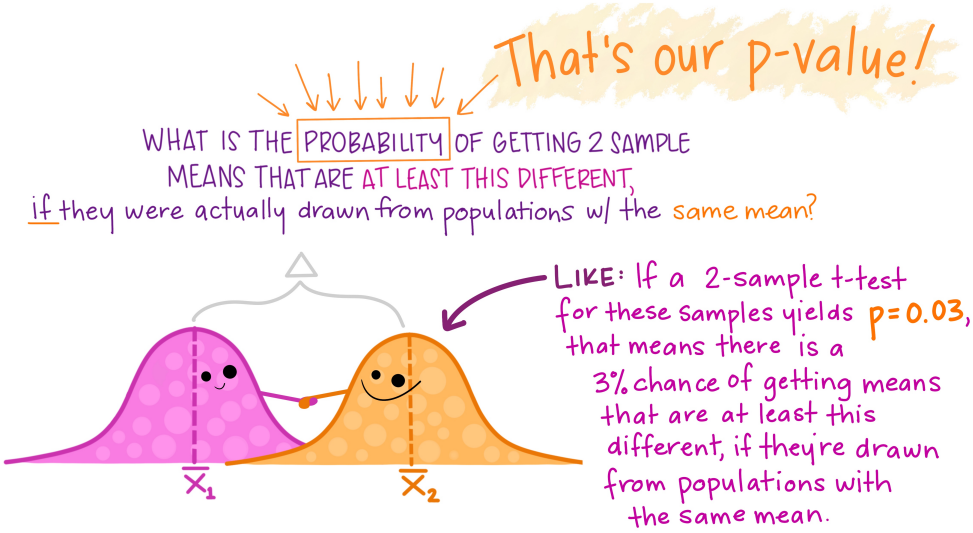
in OTHER WORDS...: The more different the sample means are,* the less likely it is they were drawn from populations w/ the same mean.

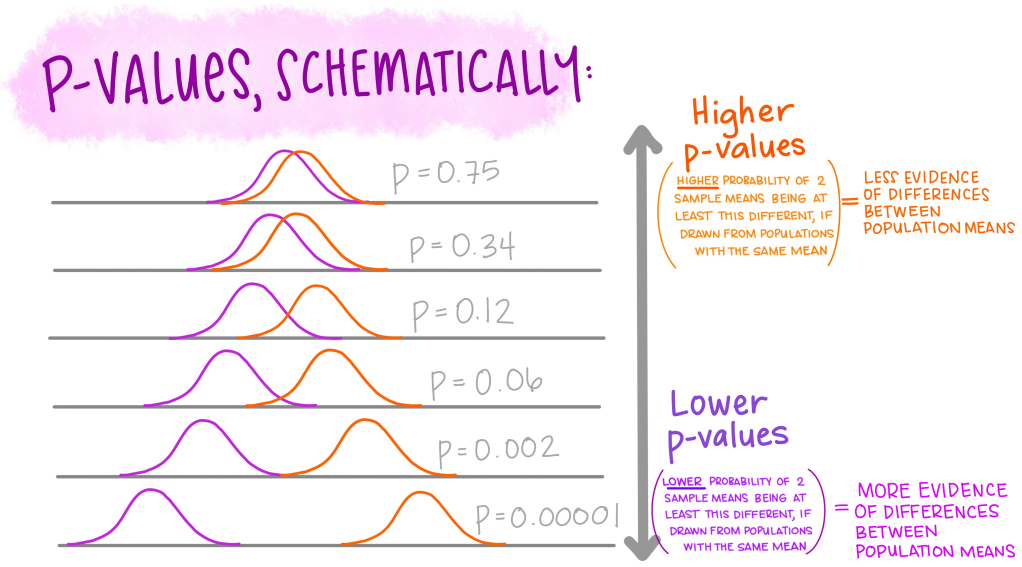
*(when taking into account sample spread + size, & assuming we've randomly sampled)



@allison_horst


So for our 2 random samples, we ask:
WHAT IS THE PROBABILITY OF GETTING 2 SAMPLE MEANS THAT ARE AT LEAST THIS DIFFERENT,*
if they were actually drawn from populations w/ the same mean?
$\bar{x}_1$
$\bar{x}_2$
*again, when taking into account sample spread & size, + assumptions...


That's our p-value!
WHAT IS THE PROBABILITY OF GETTING 2 SAMPLE MEANS THAT ARE AT LEAST THIS DIFFERENT,
if they were actually drawn from populations w/ the same mean?
LIKE: If a 2-sample t-test for these samples yields p=0.03, that means there is a 3% chance of getting means that are at least this different, if they're drawn from populations with the same mean.
$\bar{x}_1$
$\bar{x}_2$



7.4 Confidence interval for the difference between two means

In addition to reporting the p-value, we should also report a confidence interval for $\mu_X - \mu_Y$: what was the difference in means between the two groups?

> Sometimes, two groups may be statistically significantly different, but the *actual* difference may be so small as to be unimportant. Report a confidence interval as well as a p-value.

The formula for the confidence interval is

$$\bar{x} - \bar{y} \pm t_{\nu,\,0.025}\sqrt{\frac{s_X^2}{n} + \frac{s_Y^2}{m}}.$$

where the degrees of freedom ν is the same as that used in the hypothesis test. Substituting in the values, we find this confidence interval to be $[0.74, 7.8]$g. One M&M weighs about 1g, so the difference in population mean consumption might be somewhere between 1 and 8 M&Ms (at the lower limit, the effect of imagining eating M&Ms could be very small.)

7.5 Equivalence of confidence intervals and Neyman-Pearson testing

A $100(1-\alpha)\%$ confidence interval for $\mu_X = \mu_Y$ contains 0 if and only if the null hypothesis $H_0 : \mu_X = \mu_Y$ (with a two-sided alternative) is *not* rejected in a Neyman Pearson test of size α.

> If we have already calculated a $100(1-\alpha)\%$ confidence interval for $\mu_X - \mu_Y$, there is **no need** to do a separate calculation to perform a Neyman Pearson test (of size α) of the hypothesis $H_0 : \mu_X = \mu_Y$: we simply look to see whether the confidence interval contains the value 0 or not.

See the tutorial exercises for a proof.

7.6 The two-sample t test in R

In R, we use the command `t.test()`. The two samples we want to compare are stored in the vectors `treatment` and `control`, so we do

```
t.test(control, treatment)
```

```
##
##  Welch Two Sample t-test
##
## data:  control and treatment
## t = 2.4, df = 93, p-value = 0.02
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.7364 7.8263
## sample estimates:
## mean of x mean of y
##    12.388     8.106
```

We can see in the output the value of the observed test statistic `t`, the degrees of freedom in the Welch approximation `df`, the p-value, as well as a 95% confidence interval for $\mu_X - \mu_Y$.

7.6.1 t-tests with data frames in R

If your data are in a data frame, you can use a different syntax which is more convenient. Suppose the data are stored in a data frame called `eating`, with the first three and last three rows shown below

```
head(eating, n = 3)
```

```
## # A tibble: 3 x 2
##   group   amount
##   <chr>    <dbl>
## 1 Control      2
## 2 Control      8
## 3 Control     11
```

```
tail(eating, n = 3)
```

```
## # A tibble: 3 x 2
##   group     amount
##   <chr>      <dbl>
## 1 Treatment      5
## 2 Treatment      6
## 3 Treatment     12
```

The column `group` indicated which group each participant was in. We can then use the command

```
t.test(amount ~ group, data = eating)
```

```
##
##  Welch Two Sample t-test
##
## data:  amount by group
```

```
## t = 2.4, df = 93, p-value = 0.02
## alternative hypothesis: true difference in means between group Control and group Treatment i
## 95 percent confidence interval:
##  0.7364 7.8263
## sample estimates:
##   mean in group Control mean in group Treatment
##                  12.388                   8.106
```

which we can see has produced the same result. (Read the command `t.test(amount ~ group, data = eating)` as, "do a two-sample t-test to see if the mean value of the `amount` variable is different between the groups labelled by the `group` column, using the data frame `eating`").

7.7 Examples

7.7.1 Using the p-value method

Can social media be bad for your mental health and well-being? This is not a question we would expect to answer definitely with a single experiment; we would not attempt to "reject" or "not reject" a suitable hypothesis once-and-for-all. Rather, we might use hypothesis testing with the p-value method to help understand the strength of evidence provided by any single experiment.

Example 7.1 (Is quitting Facebook good for you?)**.** Tromholt (2016)[3] investigated whether quitting Facebook can improve your well-being.

In the experiment, about a thousand volunteers (all Facebook users) were randomly allocated to either a treatment group, in which they told not to use Facebook for one week, or a control group, in which they carried on using Facebook as normal. At the end of the week, all participants completed a questionnaire. One of the questions asked them to record, "In general, how satisfied are you with your life today?" on a scale of 1 (very dissatisfied) to 10 (very satisfied). Let $x_1, \ldots, x_n$ be the observed responses in the treatment group, and $y_1, \ldots, y_m$ be the observed responses in the control group. Results from those who repsponded were as follows.

$$\bar{x} = 8.11,\ \bar{y} = 7.74,\ s_X^2 = 1.23^2,\ s_Y^2 = 1.43^2,\ n = 516,\ m = 372$$

with

$$\nu = \frac{\left(\frac{s_X^2}{516} + \frac{s_Y^2}{372}\right)^2}{\frac{(s_X^2/516)^2}{516-1} + \frac{(s_Y^2/372)^2}{372-1}} \simeq 726.$$

1. Defining your notation carefully, state suitable hypotheses for the experiment.
2. Conduct an appropriate hypothesis test, reporting the p-value.
3. Report a 95% confidence interval for the difference in population means.
4. In plain English, summarise your results.

Some R output to help is as follows:

[3] Morten Tromholt Cyberpsychology, Behavior, and Social Networking 2016 19:11, 661-666.

```
pt(-4.03, 726)
```

```
## [1] 3.083e-05
```

```
qt(0.975, 726)
```

```
## [1] 1.963
```

Solution

1. Define μ_X to be the population mean "life satisfaction score" under the treatment group condition, and μ_Y to be the population mean score under the control group condition. Our hypotheses are

$$H_0 : \mu_X = \mu_Y,$$
$$H_A : \mu_X \neq \mu_Y.$$

2. Using the two-sample t-test, we compute

$$t_{obs} = \frac{\bar{x} - \bar{y}}{\sqrt{\frac{s_x^2}{516} + \frac{s_y^2}{372}}} = 4.03.$$

   For the degrees of freedom parameter, we are given $\nu \simeq 726$, so under H_0, the test statistic would have (approximately) a t_{726} distribution.. The observed test statistic is right out in the tail of this distribution, so the p-value will be small.

The p-value is given by

$$2 \times P(T_{726} \leq -4.03) \simeq 6 \times 10^{-5}$$

3. An approximate 95% confidence interval for the difference in population means is

$$\bar{x} - \bar{y} \pm t_{726,\,0.025} \sqrt{\frac{s_X^2}{n} + \frac{s_Y^2}{m}}.$$

   From the R output, we have $t_{726,\,0.025} = 1.963$, so the confidence interval is $[0.2, 0.6]$, to 1 decimal place.

4. The experiment found very strong evidence against the null hypothesis of equal mean life satisfaction scores, with the particants who did not use Facebook for one week giving higher scores. However, the effect of not using Facebook was small: the difference in means is likely less than a single point on the 10 point response scale.

7.7.2 Using Neyman-Pearson testing

Neyman-Pearson testing is used is in medical research, specifically, clinical trials for new drugs. The scenario would be something like this:

- A pharmaceutical company has developed a new drug, and will test it using a hypothesis test.
- The null hypothesis is that the drug has no effect.
- The action to be taken following the test is either to license the drug for use on patients, or to

decide that it cannot be used; the pharmaceutical company would then abandon that drug, and move onto to developing a different drug.
- If the null hypothesis is "rejected", we conclude that the drug *does* has an effect, and the drug gets its license (assuming the drug effect is beneficial for patients).
- If the null hypothesis is "not rejected", we conclude that there is no evidence the drug works, and it is not licensed for further use.

Example 7.2 (Testing a new diabetes treatment.)**.** Patients with type-2 diabetes may use drugs to control their blood sugar levels. A pharmaceutical company (Merck) conducted a clinical trial to compare the efficacy of a combination of two drugs, sitagliptin and metformin, with using metformin alone. The product name for this combination of drugs is "Efficib". 190 patients were recruited to the trial, and were randomly allocated to one of two treatments:

- treatment 1: 100mg sitagliptin per day, and at least 1500mg metformin per day
- treatment 2: a daily placebo, made to look like a dose of 100mg sitagliptin, and at least 1500mg metformin per day.

The study was "double-blinded": neither the patients nor their doctors knew which treatment they were getting (though the trial investigators did know.) A1C (a measure of blood sugar level) was recorded for each patient at the start and after 18 weeks, and the change in A1C was recorded for each patient.

We model this as follows.

Let X_i denote the change in A1C for the i-th patient on the treatment 1, and Y_i denote the change in A1C the i-th patient on treatment 2. We suppose

$$X_1, \dots, X_{95} \overset{i.i.d}{\sim} N(\mu_X, \sigma_X^2),$$
$$Y_1, \dots, Y_{92} \overset{i.i.d}{\sim} N(\mu_Y, \sigma_Y^2).$$

We denote the corresponding observed values by $x_1, \dots, x_{95}$ and $y_1, \dots, y_{92}$. The trial results are published at clinicaltrials.gov. Individual patient data are not normally published, and we can infer (approximately) what the summary statistics were: we have

$$\bar{x} = \frac{1}{95}\sum_{i=1}^{95} x_i = -1.00, \tag{7.5}$$

$$s_X^2 = \frac{1}{94}\sum_{i=1}^{95}(x_i - \bar{x})^2 = 1.5456, \tag{7.6}$$

$$\bar{y} = \frac{1}{92}\sum_{i=1}^{92} y_i = 0.02, \tag{7.7}$$

$$s_Y^2 = \frac{1}{91}\sum_{i=1}^{92}(y_i - \bar{y})^2 = 1.4968. \tag{7.8}$$

1. State appropriate null and hypotheses, in terms of your model parameters, to test whether the addition of sitagliptin had an effect
2. Conduct an appropriate Neyman-Pearson test, of size 0.05, stating the conclusions clearly.

Some R output to help is as follows.

```
qt(c(0.9, 0.95, 0.975, 0.99), 185)
```

```
## [1] 1.286 1.653 1.973 2.347
```

Solution

We consider the hypotheses

$$H_0 : \mu_X = \mu_Y,$$
$$H_A : \mu_X \neq \mu_Y,$$

so that the null hypothesis is that there is no effect from the additional treatment with sitagliptin

For our two-sample t-test, we have

$$t_{obs} = \frac{\bar{x} - \bar{y}}{\sqrt{\frac{s_X^2}{95} + \frac{s_Y^2}{92}}} = -5.655,$$

$$\nu = \frac{\left(\frac{s_X^2}{95} + \frac{s_Y^2}{92}\right)^2}{\frac{(s_X^2/95)^2}{94} + \frac{(s_Y^2/92)^2}{91}} \simeq 185.$$

Hence, for a test of size 0.05, our critical region is anything above $t_{185;\,0.025}$ or anything below $t_{185;\,0.975}$ (anything above the 97.5th percentile, or anything below the 2.5th percentile, for the student-t distribution with 185 degrees of freedom.) From the R output, we have $t_{17.7;\,0.025} = 1.97$ and so $t_{17.7;\,0.975} = -1.97$. The critical region and observed test statistic are plotted below.

As t_{obs} does lie in the critical region, we conclude that we reject H_0. We say that there is evidence (at the 5% level of significance) that there is an effect of combining sitagliptin with metformin, and that this effect is an increased reduction in A1c (adding sitagliptin has a beneficial effect.)

There have been other studies to test the effect of sitagliptin and metformin (the "Efficib" drug). Based on these studies, the European Medicines Agency approved Efficib for use in the European Union.

Chapter 8

Hypothesis testing: comparing two proportions

In this chapter we will test whether the probability parameters θ_X and θ_Y in two binomial distributions $X \sim Bin(n, \theta_X)$ and $Y \sim Bin(m, \theta_Y)$, are equal or not, given observations from each distribution. In particular, when might we conclude that θ_X and θ_Y are different, based on an observed difference between X/n and Y/m?

We will again use two different methods: a computer simulation method, and an analytical method based on the normal distribution.

8.1 Example: an investigation into gender bias

Steinpreis et al. (1999)[1] conducted the following experiment. CVs were sent to male and female academic psychologists at various US universities. The psychologists were asked whether or not they would hire the applicant for an academic job based on the CV. The CVs sent to the psychologists were identical *except* for the name of the applicant: "Brian Miller" on some, and "Karen Miller" on the others. The interest was in whether the gender of the applicant made a difference: whether male applicants were more or less likely to be hired than female applicants.

Results from the experiment were as follows. The "recruiters" are the academic psychologists. There are 128 different recruiters: each recruiter sees one CV only, where the applicant is either male or female.

applicant	recruiter	hired	rejected	% hired
male	male	24	7	77.4%
female	male	16	16	50.0%
male	female	22	10	68.8%
female	female	13	20	39.4%

The data suggest a clear gender bias: male applicants are more likely to be hired (regardless of the gender of the recruiter). But can we be sure of this? We'd expect recruiters to have different opinions

[1] Steinpreis, R. E., Anders, K. A. and Rizke, D. (1999). The Impact of Gender on the Review of the Curricula Vitae of Job Applicants and Tenure Candidates: A National Empirical Study. Sex Roles, Vol. 41, Nos. 7/8.

anyway, and we can see that within each row of the table, some recruiters must have been more demanding of their applicants than others, in that some chose to hire, and others chose to reject. Perhaps we were just unlucky with our sample of recruiters? For example, perhaps in row two of the table, recruiters tended to be more demanding than the recruiters in row one? We can investigate this using a hypothesis test (as did the study authors, though they used a slightly different method.)

8.2 Comparing two binomial proportions

To simplify things, we'll just consider the male recruiters:

applicant	recruiter	hired	rejected	% hired
male	male	24	7	77.4%
female	male	16	16	50.0%

The observed difference in % hired for the two groups was 27.4% Could a difference this large arise purely by chance?

We use a binomial model for the data, with a separate binomial distribution for the number of recruiters choosing to hire in each row of the table: defining X as the number of recruiters who would hire the male applicant, and Y as the number of recruiters who would hire the female applicant, we suppose that

$$X \sim Binomial(n, \theta_X), \tag{8.1}$$
$$Y \sim Binomial(m, \theta_Y), \tag{8.2}$$

with $n = 31$ and $m = 32$. We interpret θ_X and θ_Y as, respectively, the proportion of all recruiters in the population who would hire the male applicant, and the proportion of all recruiters in the population who would hire the female applicant.

If the gender of the applicant was irrelevant to all recruiters, then we would have $\theta_X = \theta_Y$, and we will write our null hypothesis as

$$H_0 : \theta_X = \theta_Y.$$

8.2.1 A simulation method

As before, we need to understand what sort of data could arise *purely by chance*. In our gender bias example, we need to understand how different X/n and Y/m could be, if H_0 were true and the probabilities θ_X and θ_Y were the same.

In the experiment, the observed values of X and Y were 24 and 16 (with the observed difference in proportions being $\frac{24}{31} - \frac{16}{32} \simeq 27\%$).

- If it's (almost) impossible to get a difference this large purely by random chance, we would conclude that the experiment has provided evidence *against* the hypothesis that the two probabilities θ_X and θ_Y are equal.
- If it's easy to get a difference this large purely by random chance, we *won't* say this shows H_0 is true, but we *will* say that the experiment has *failed to provide evidence against* H_0.

Let's now see what can happen purely by chance, using simulation. We will need to choose θ_X and θ_Y, which we need to be equal if we are assuming H_0 is true. We'll choose these probabilities to equal the total number of hires (40) divided by the total number of recruiters (63).

We'll first simulate five X, Y pairs in R: we simulate five observations from the $Binomial(31, 40/63)$ distribution, and store the result in the vector `males` and five observations from the $Binomial(32, 40/63)$ distribution, and store the result in the vector `females`:

```
males <- rbinom(n = 5, size = 31, prob = 40/63)
females <- rbinom(n = 5, size = 32, prob = 40/63)
```

Then to see what we've got:

```
males
```

```
## [1] 21 21 19 16 22
```

```
females
```

```
## [1] 17 16 19 19 24
```

and to compare the proportions:

```
males/31 - females/32
```

```
## [1]  0.14617  0.17742  0.01915 -0.07762 -0.04032
```

The first pair generated for (X, Y) was (21, 17), and the difference between the two proportions was $\frac{21}{31} - \frac{17}{32} \simeq 0.15$: we got a 15% difference in the proportions hired, just by random chance. However, we didn't get anything as large as the *observed* difference of 27%. Now we will do this a large number of times, and look at the distribution of the difference between the proportions:

```
males <- rbinom(n = 100000, size = 31, prob = 40/63)
females <- rbinom(n = 100000, size = 32, prob = 40/63)
differences <- males/31 - females/32
```

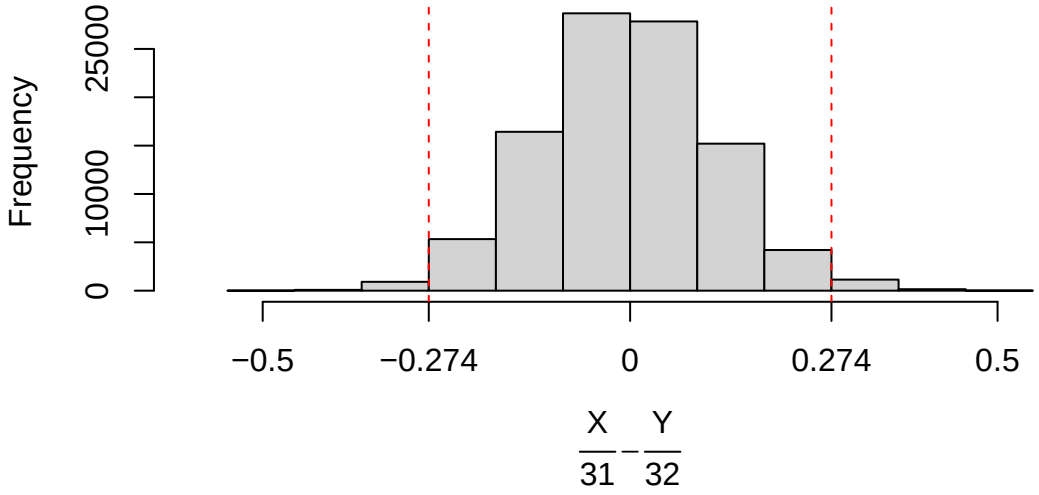


Figure 8.1: Histogram of simulated differences $X/31 - Y/32$, randomly generated assuming H_0 is true. It is possible to obtain differences larger than 0.274 (the difference observed in the experiment), but not very likely; it's hard to obtain a difference this large by random chance alone.

We can see that it is possible to get a difference as large as 0.274, but not that likely. Out of the 100,000 simulations, we can count how many times this happened:

```
sum(differences >= 0.274)
```

```
## [1] 1306
```

so we would estimate the probability of seeing a difference as large as 0.274, purely by random chance, to be 1306 / 100000 $\simeq$ 0.013.

What about the negative differences, in particular those, below -0.274? Should we count those? This depends on whether we want to report

1. how *far apart* X/n and Y/m could be, purely by random chance, or
2. how *much greater* X/n could be than Y/m, purely by random chance.

In this case, a large negative value of $X/n - Y/m$ would still suggest unequal treatment of males and females, so we should report the first case above. This means that we are using a **two-sided** alternative hypothesis

$$H_A : \theta_X \neq \theta_Y,$$

rather than a **one-sided** alternative $H_A : \theta_X > \theta_Y$.

So we calculate how many times we generated obtained $|X/31 - Y/32| \geq 0.274$ (use the `abs()` command in R to get the absolute value)

```
sum(abs(differences) >= 0.274)
```

```
## [1] 2308
```

So in conclusion, we report the value of

$$P\left(\left|\frac{X}{31}-\frac{Y}{32}\right|\geq 0.274\right),$$

assuming H_0 to be true, which we estimate to be 100000.274 / 100000 $\simeq$ 0.023.

In summary:

we estimate about a 2% probability that, by nothing other than random chance, the (absolute) difference in percentages of hired applicants between the genders could be as large as 27.4% (the difference that was observed in the experiment).

This probability of 2% is a p-value: a probability of getting a difference as extreme as the one we observed, assuming H_0 to be true.

8.3 An analytical method

So we can write this in general terms, we will define x and y as the observed values of the random variables X and Y (in the example, we have $x = 21$ and $y = 17$). We used simulation to estimate

$$P\left(\left|\frac{X}{n}-\frac{Y}{m}\right|\geq\left|\frac{x}{n}-\frac{y}{m}\right|\right),$$

assuming H_0 to be true. We will now attempt to work out this probability analytically, by expressing it in terms of a standard probability distribution. We have

$$P\left(\left|\frac{X}{n}-\frac{Y}{m}\right|\geq\left|\frac{x}{n}-\frac{y}{m}\right|\right)=P\left(\frac{\left|\frac{X}{n}-\frac{Y}{m}\right|}{\sqrt{v}}\geq\frac{\left|\frac{x}{n}-\frac{y}{m}\right|}{\sqrt{v}}\right)$$

where we define

$$v:=p^*(1-p^*)\left(\frac{1}{n}+\frac{1}{m}\right)$$

with

$$p^*:=\frac{x+y}{n+m}$$

Now we define

$$Z:=\frac{\frac{X}{n}-\frac{Y}{m}}{\sqrt{v}}$$

If we assume H_0 is true, and we further assume $\theta_X=\theta_Y=\frac{x+y}{n+m}$ (just as we did to simulate our random data), then we have

$$\mathbb{E}(Z)=0, \tag{8.3}$$

$$Var(Z)=1. \tag{8.4}$$

(Deriving these results is an exercise in the tutorial questions.)

We now make the *approximation* that $Z\sim N(0,1)$ (using the result that a $Binomial(n,p)$ distribution can be approximated by a $N(np,np(1-p))$ distribution, for 'large' n and 'moderate' p.).

We *didn't* have to make any approximations about normal distributions in the simulation method, so we can think of that as more 'accurate' than this analytical method. But maybe we'll get similar results! We will soon see...

To compute the p-value, we can write

$$P\left(\left|\frac{X}{n}-\frac{Y}{m}\right| \geq \left|\frac{x}{n}-\frac{y}{m}\right|\right) = P\left(|Z| \geq \frac{\left|\frac{x}{n}-\frac{y}{m}\right|}{\sqrt{v}}\right) \tag{8.5}$$

$$P\left(Z \leq -\frac{\left|\frac{x}{n}-\frac{y}{m}\right|}{\sqrt{v}}\right) + P\left(Z \geq \frac{\left|\frac{x}{n}-\frac{y}{m}\right|}{\sqrt{v}}\right) \tag{8.6}$$

$$\simeq \Phi\left(-\frac{\left|\frac{x}{n}-\frac{y}{m}\right|}{\sqrt{v}}\right) + 1 - \Phi\left(\frac{\left|\frac{x}{n}-\frac{y}{m}\right|}{\sqrt{v}}\right), \tag{8.7}$$

where $\Phi(.)$ is the cumulative distribution function of the $N(0,1)$ distribution.

In our example, with $n = 31, m = 32, x = 24, y = 16$, we compute

$$p^* = \frac{24+16}{31+32}, \quad v = p^*(1-p^*)\left(\frac{1}{31}+\frac{1}{32}\right)$$

and our p-value is

$$\Phi(-2.2599) + 1 - \Phi(2.2599) = 0.024,$$

to 3 d.p. Notice how similar this is to the p-value computed using simulation. We visualise the p-value in Figure 8.2, where we plot the distribution of Z under H_0. For comparison, we also plot the histogram of our simulated values $X/31 - Y/32$, each now divided by $\sqrt{v}$.

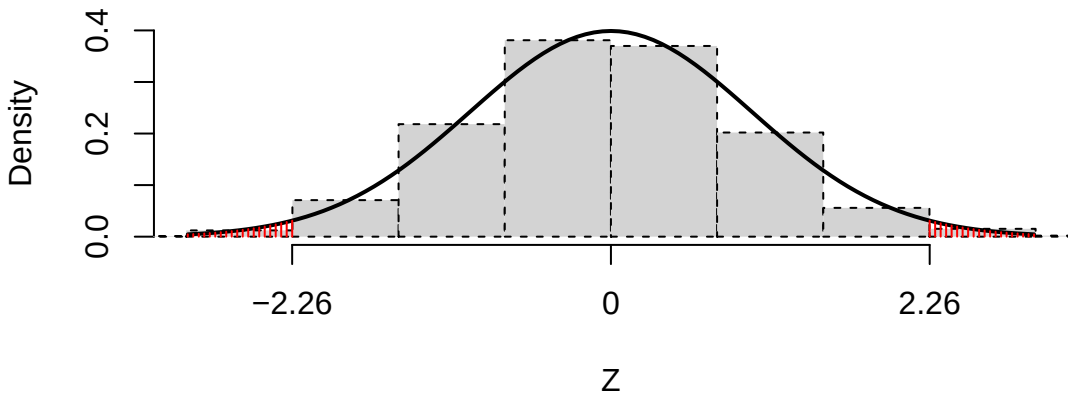


Figure 8.2: The distribution of the test statistic Z under H_0. For comparison, a histogram shows the distribution of test statistics obtained using random simulation: note the close agreement.

8.3.1 The analytical method: a summary

To summarise, the steps are as follows.

1. State the model and hypotheses

We wish to compare two binomial proportions. We have

$$X \sim Bin(n, \theta_X),$$
$$Y \sim Bin(m, \theta_Y)$$

and our null hypothesis is

$$H_0 : \theta_X = \theta_Y,$$

the 'success' probabilities in our two samples are the same. For a two-sided alternative, we have

$$H_A : \theta_X \neq \theta_Y$$

2. **Choose an appropriate test statistic**

The test statistic measures the difference between the two sample proportions. We use use the test statistic

$$Z = \frac{\frac{X}{n} - \frac{Y}{m}}{\sqrt{P^*(1-P^*)\left(\frac{1}{n} + \frac{1}{m}\right)}}, \tag{8.8}$$

where

$$P^* = \frac{X+Y}{n+m} \tag{8.9}$$

3. **State the distribution of the test statistic, under the assumption that H_0 is true**

We think of Z as a random variable, because it is a function of the two binomial random variables X and Y. If H_0 is true, then approximately, we have

$$Z \sim N(0, 1).$$

4. **Calculate the test statistic for the observed data**

Remembering that we denote the values we actually observed by x and y, the corresponding observed value of the test statistic is

$$z_{obs} = \frac{\frac{x}{n} - \frac{y}{m}}{\sqrt{p^*(1-p^*)\left(\frac{1}{n} + \frac{1}{m}\right)}},$$

where

$$p^* = \frac{x+y}{n+m}.$$

5. **Report the evidence against the null hypothesis, by calculating the p-value**

We have the same definition of the p-value as before, but now using the test statistic Z and its corresponding distribution under H_0:

$$P(|Z| \geq |z_{obs}|),$$

8.3.2 Conclusion

In statistical terms, we would say that with a p-value between 0.01 and 0.05, we have 'weak' evidence against the null hypothesis. Given such a p-value (and noting that the experiment was fairly small in any case), it would be desirable to replicate the experiment, to see if the results are the same. This has indeed happened: see for example Moss-Racusin et al. (2012)[2]], who conducted a similar study, and observed a similar bias against female job applicants.

8.4 Confidence intervals to measure the difference

An approximate 95% confidence interval for the difference $\theta_X - \theta_Y$ is given by

$$\frac{x}{n} - \frac{y}{m} \pm 1.96\sqrt{v}, \tag{8.10}$$

with

$$v = p^*(1-p^*)\left(\frac{1}{n} + \frac{1}{m}\right), \quad p^* = \frac{x+y}{n+m}.$$

In our example, we obtain a 95% confidence interval of (3.6%, 51.2%): this is wide in this context, reflecting a lot of uncertainty.

Example 8.1 (Hypothesis testing: comparing binomial proportions. Can early release and tagging of prisoners affect the likelihood of reoffending?)**.** Meuer and Woessner (2018)[3] describe an experiment to test the effect of electronic monitoring (tagging) on "low-risk" prisoners. We describe some of their data here. Forty-eight (male) prisoners were randomly allocated to two groups:

- in the experimental group, the prisoner served the last part of his sentence under "supervised early work release", involving the use of an open prison and electronic tagging.
- in the control group, the prisoner served the last part of his sentence in prison, as normal.

Following the end of the sentence, the prisoners were followed up for two years. It was recorded whether each prisoner reoffended. The results were as follows.

group	sample size	number reoffending	% reoffending
experimental	24	7	29.2%
control	30	15	50.0%

1. Specify an appropriate model for these data and hypothesis to test.
2. Use the following R output to assess whether there is evidence that early work release/tagging scheme has affected the probability of reoffending

```
experimental <- rbinom(n = 100000, size = 24, prob = 22 / 54)
control <- rbinom(n = 100000, size = 30, prob = 22 / 54)
differences <- experimental / 24 - control / 30
sum(abs(differences) >= abs(7/24 - 15/30))
```

[2]Faculty's subtle gender biases favor male students, Corinne A. Moss-Racusin, John F. Dovidio, Victoria L. Brescoll, Mark J. Graham, Jo Handelsman, Proceedings of the National Academy of Sciences Sep 2012, 201211286; DOI: 10.1073/pnas.1211286109

[3]Meuer, K. and Woessner, G. (2018). Does electronic monitoring as a means of release preparation reduce subsequent recidivism? A randomized controlled trial in Germany. European Journal of Criminology, 1-22.

```
## [1] 13026
```

3. Conduct a suitable hypothesis test using the normal approximation. Draw a sketch that indicates the p-value. Based on the output above, what do you think the p-value would be?

4. Calculate a 95% confidence interval for the difference between the two probabilities of reoffending.

Solution

1. Define the random variables X and Y as the numbers reoffending in the experimental and control groups respectively. We suppose

$$X \sim Bin(24, \theta_X), \quad Y \sim Bin(30, \theta_Y), .$$

and our hypotheses are

$$H_0 : \theta_X = \theta_Y, \quad H_A : \theta_X \neq \theta_Y.$$

2. The observed difference in proportion was

$$\frac{7}{24} - \frac{15}{30} = -0.208,$$

(to 3 d.p.). In the R code, we have simulated X, Y pairs assuming H_0 is true, with $\theta_X = \theta_Y = \frac{7+15}{24+30}$. The last line counts how many times we simulated a pair where the (absolute) difference in proportions was at least 0.208. This happened 13026 times out of 100,000, so we estimate that there is 13% probability of observing, purely by random chance, a difference as large at that seen in the experiment. This probability is relatively high, giving no evidence against H_0: no evidence that the early work release/tagging scheme has affected the probability of reoffending.

3. We compute

$$z_{obs} = \frac{\frac{7}{24} - \frac{15}{30}}{\sqrt{p^*(1-p^*)\left(\frac{1}{24} + \frac{1}{30}\right)}} = -1.54,$$

(with $p^* = (7 + 15)/(24 + 30)$)

The p-value is obtained from

$$P(|Z| \geq 1.54),$$

where $Z \sim N(0, 1)$, and is shown as the shaded area in the following plot.

from the R output in part (2), we would expect this to be around 0.13. We can obtain the p-value from R as follows.

4. An approximate 95% confidence interval for the difference in proportions is

$$-0.208 \pm 1.96\sqrt{v},$$

with $v = p^*(1 - p^*)(1/24 + 1/30)$ and $p^* = (7 + 15)/(24 + 30)$, which gives

$$(-47\%, 5.5\%)$$

Chapter 9

Sample size and power for a Neyman-Pearson hypothesis test

How do we choose a sample size when designing a hypothesis testing experiment? Suppose we plan to use the Neyman-Pearson framework (so the conclusion will be that we either "reject" or "do not reject" the null hypothesis). Recall the two types of mistake we can make: a Type I error, of incorrectly rejecting H_0 when H_0 is true, and a Type II error: failing to reject H_0 when H_0 is false.

- We choose (in advance) the size/significance level of the test: this determines the Type I error rate.
- Our choice of sample size will influence the Type II error rate: the larger the sample size, the smaller the probability of a Type II error.

9.1 Gender bias example re-visited

Consider again the gender bias experiment from the previous chapter. Suppose we want to repeat it with new data, with 30 recruiters in each group, to see if we can confirm the original findings. With the same notation as before, we have

$$X \sim Bin(n = 30, \theta_X), \tag{9.1}$$
$$Y \sim Bin(m = 30, \theta_Y), \tag{9.2}$$

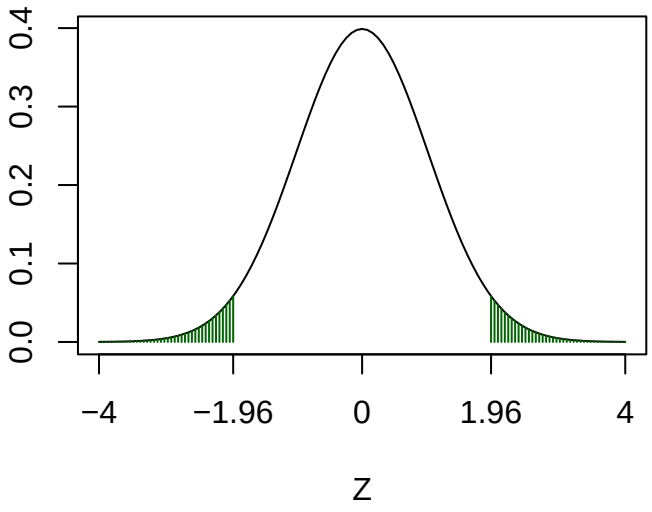
and the hypotheses

$$H_0 : \theta_X = \theta_Y,$$
$$H_A : \theta_X \neq \theta_Y,$$

Suppose we are going to use a Neyman-Pearson test of size 0.05: we will either "reject" or "not reject" H_0. Our test statistic is

$$Z := \frac{\frac{X}{n} - \frac{Y}{m}}{\sqrt{v}}$$

and assuming H_0 is true, we have Z $N(0,1)$, so the critical region will be $(-\infty, -1.96) \cup (1.96, \infty)$:



If H_0 is true, then the probability that our test statistic Z will fall in the critical region is exactly 0.05. But what if H_0 is false, and we actually *want* Z to fall in the critical region? What is the probability this would happen? If H_0 is false, then the probability of (correctly) rejecting H_0 will depend on

1. the difference between θ_X and θ_Y;
2. the sample sizes n and m.

We'll investigate this with a simulation. Let's suppose we have $\theta_X = 0.75$ and $\theta_Y = 0.5$, and $n = m = 30$, so that, in this case, we would have

$$X \sim Bin(30, 0.75), \quad Y \sim Bin(30, 0.5).$$

We first sample a random X and Y

```
x <- rbinom(1, size = 30, prob = 0.75)
y <- rbinom(1, size = 30, prob = 0.5)
c(x, y)
```

```
## [1] 25 20
```

So in this simulation our observed values of X and Y are 25 and 20. We now compute our observed test statistic:

```
pstar <- (x + y) / (30 + 30)
v <- pstar * (1 - pstar) * (1 / 30 + 1 / 30)
z <- (x/30 - y/30)/sqrt(v)
z
```

```
## [1] 1.491
```

Our observed test statistic would be 1.49, which is not in the critical region, so H_0 would *not* be rejected in this case. Now let's repeat this lots of times, and see how often H_0 would be rejected

```
x <- rbinom(100000, size = 30, prob = 0.75)
y <- rbinom(100000, size = 30, prob = 0.5)
pstar <- (x + y) / (30 + 30)
v <- pstar * (1 - pstar) * (1 / 30 + 1 / 30)
z <- (x/30 - y/30)/sqrt(v)
sum(abs(z) > 1.96)
```

```
## [1] 51585
```

So, there's about a 52% chance (51585/100000) we'd get data resulting in our test statistic falling in the critical region. We visualise this below.

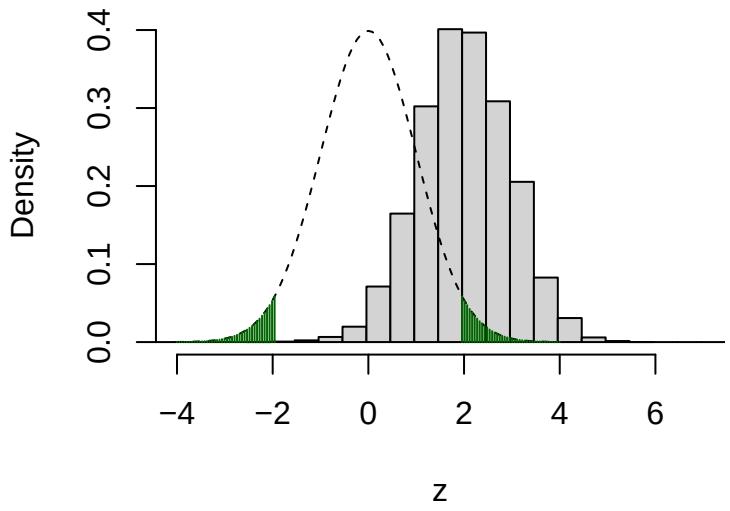


Figure 9.1: The dashed line shows the assumed distribution of the test statistic Z under H_0, and the critical region (shaded green). The histogram shows the *real* distribution of Z, which we have obtained by simulation. About 52% of the mass of this distribution lies in the upper critical region (1.96 and above), so there's about a 52% chance of the test statistic falling where we want it to, in this scenario where H_0 really is false.

What happens if we increase the sample size, say to $n = m = 100$?

```
x <- rbinom(100000, size = 100, prob = 0.75)
y <- rbinom(100000, size = 100, prob = 0.5)
pstar <- (x + y) / (100 + 100)
v <- pstar * (1 - pstar) * (1 / 100 + 1 / 100)
z <- (x/100 - y/100)/sqrt(v)
```

```
## [1] 95989
```

Now, there's about a 96% chance (95989/100000) we'd get data resulting in our test statistic falling in the critical region. We visualise this below.

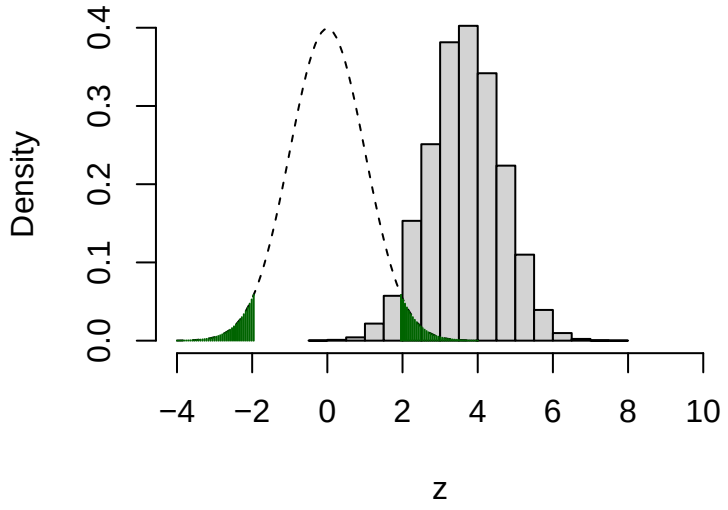


Figure 9.2: The dashed line shows the assumed distribution of the test statistic Z under H_0, and the critical region (shaded green). The histogram shows the *real* distribution of Z, which we have obtained by simulation. About 96% of the mass of this distribution lies in the upper critical region (1.96 and above), so there's about a 96% chance of the test statistic falling where we want it to, in this scenario where H_0 really is false.

9.2 The power of a hypothesis test

Definition 9.1 (Power)**.** The power of a hypothesis test is defined to be the probability that we will reject H_0 when H_0 is false. This is the probability of *avoiding* a Type II error.

The definition above is a bit vague, in that the probability will depend on what the various population parameters are (the values of the binomial probabilities in our example); it will depend on *how* exactly H_0 is false. To obtain a power, we will need to specify values for all the quantities that determine the distribution of the test statistic, including the sample size.

In the example above, we found the following.

- For $\theta_X = 0.75$, $\theta_Y = 0.5$, and $n = m = 30$, the power was estimated to be 0.52: there was a 52% chance of rejecting H_0.
- For $\theta_X = 0.75$, $\theta_Y = 0.5$, and $n = m = 100$, the power was estimated to be 0.96: there was a 96% chance of rejecting H_0.

We can see that increasing the sample size increases the power. The power will also increase as $|\theta_X - \theta_Y|$ increases: as the difference between the two probability parameters increases, it becomes easier to detect that the two population proportions are different.

9.3 An analytical approach

As before, we can use an analytical approach rather than simulation. We just state the result here, and leave the derivation as an exercise in the tutorial questions. In the case of equal sample sizes per

group ($n = m$), the power is given by $\Phi(Z^*)$, where $\Phi(.)$ is the cumulative distribution function of the standard normal distribution, and

$$Z^* = \sqrt{\frac{n(\theta_X - \theta_Y)^2}{\theta_X(1-\theta_X) + \theta_Y(1-\theta_Y)}} - Z_{\alpha/2} \tag{9.3}$$

(where, for example, if $\alpha = 0.05$ then $Z_{\alpha/2} = 1.96$)

In our two examples, this gives a power of 0.97 for $n = m = 100$ and 0.54 for $n = m = 30$; this agrees well with the simulations.

Formulae such as (9.3) give us a means of choosing a sample size for an experiment. We decide what power we want (80% is a common choice), make an assumption about the population parameters (θ_X and θ_Y), and then work out what sample size is needed to achieve the desired power.

Example 9.1 (Sample size and power calculation for a hypothesis test.)**.** Consider again the electronic tagging example. Using the same notation suppose the true probabilities are $\theta_X = 0.3$ and $\theta_Y = 0.5$. The experiment is to be repeated, with a test of size 0.05.

1. If the sample sizes were $n = m = 20$, what does the following R code suggest would be the power?

```
x <- rbinom(100000, size = 20, prob = 0.3)
y <- rbinom(100000, size = 20, prob = 0.5)
pstar <- (x + y) / (20 + 20)
v <- pstar * (1 - pstar) * (1 / 20 + 1 / 20)
z <- (x/20 - y/20)/sqrt(v)
sum(abs(z) > 1.96)
```

```
## [1] 24806
```

2. Assuming an equal sample size in each group, what sample size would be needed to achieve 90% power? Use the following R output to help you.

```
qnorm(0.9)
```

```
## [1] 1.282
```

Solution

1. In the R code, 100,000 test statistics were simulated, with sample sizes of $n = m = 20$ and probability parameters $\theta_X = 0.3$ and $\theta_Y = 0.5$. We observe 24304 lying in the critical region, so the power of the test would be estimated as $24303/100000 \simeq 0.24$: only be a 24% chance of correctly rejecting H_0.
2. We require $\Phi(Z*) = 0.9$, so we have $Z^* = \Phi^{-1}(0.9)$. The R output gives us $\Phi^{-1}(0.9) = 1.282$ (to 3 d.p.). Now we invert equation (9.3), to get

$$n = (1.96 + 1.282)^2 \times \frac{0.5^2 + 0.3 \times 0.7}{0.2^2} \simeq 121$$

so we would need about 121 people in each group.

Chapter 10

χ^2 tests for contingency tables

For our final hypothesis testing problem, we will extend the problem of comparing two binomial distributions to two (or more) multinomial distributions, to see if the probability parameters in each distribution are the same.

10.1 Example: customer ratings of restaurants

The website tripadvisor gives customer reviews of hotels, restaurants, tourist attractions etc., and each review includes a rating. Ratings for restaurants in Sheffield (in September 2018) were as follows:

This type of data is sometimes described as **contingency table** data. For each customer, we record two qualitative variables: the restaurant the customer went to, and the rating the customer gave. The numbers in the table then give the number of times each restaurant-rating pair occurred.

Based on these scores, the restaurants were given a ranking: Akbar's was ranked 187 out of 1257 restaurants in Sheffield, and Aagrah was ranked 116. However, the percentage of ratings in each category looks similar for the two restaurants, as we can see in Figure 10.1. We might wonder whether the customer ratings are significantly different; if they are not, one could argue that the rankings are not meaningful. (The same concern applies to most league tables, sporting ones excluded.)

Table 10.1: Ratings for two restaurants on Tripadvisor (as of 12th September, 2018). There were 298 reviews for Akbar's, and 916 reviews for Aagrah.

	Excellent	Very good	Average	Poor	Terrible
Akbar's	146	70	33	24	25
Aagrah	419	277	102	66	52

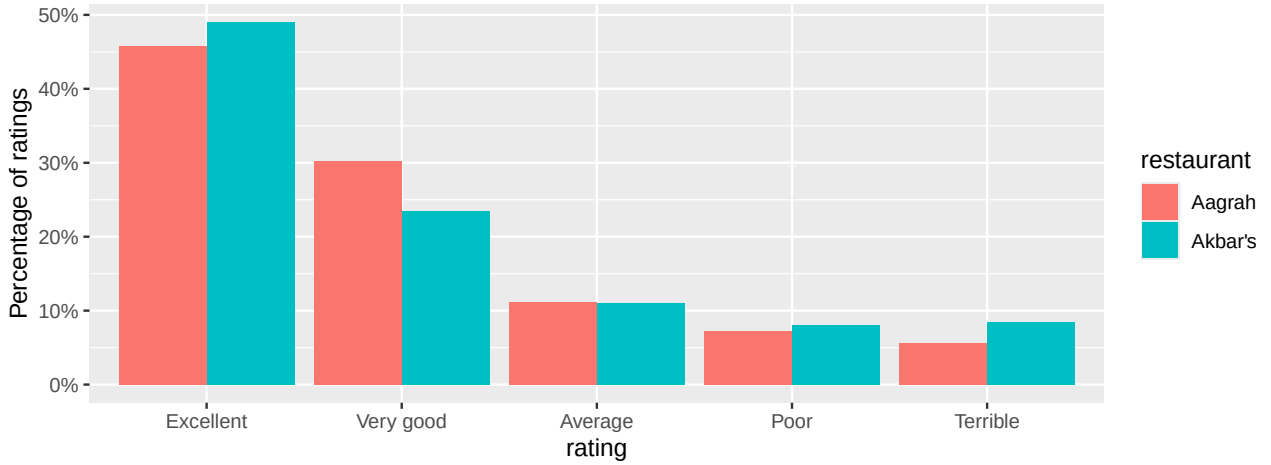


Figure 10.1: Comparing percentages of ratings for the two restaurants. These look similar - could the differences just be down to random chance?

10.2 A model and hypotheses

We can model these data using multinomial distributions. Before the ratings have been observed, define, as random variables, $Y = (Y_1, Y_2, Y_3, Y_4, Y_5)$ as the number of ratings in each category in for Akbar's, and $Z = (Z_1, Z_2, Z_3, Z_4, Z_5)$ as the number of ratings in each category for Aagrah So, for example, Y_2 would be the number of "very good" ratings for Akbar's, and Z_5 would be the number of "terrible" ratings for Aagrah. Keeping the total number of ratings fixed for each restaurant, we might suppose that

$$Y \sim multinom(298; \theta_1, \theta_2, \ldots, \theta_5), \tag{10.1}$$
$$Z \sim multinom(916; \phi_1, \phi_2, \ldots, \phi_5), \tag{10.2}$$

with, for example, θ_1 the probability of an "excellent" rating for Akbar's, and ϕ_2 the probability of a "very good" rating for Aagrah.

We state the null hypothesis as

$$H_0 : \theta_i = \phi_i, \quad i = 1, 2, \ldots, 5 \tag{10.3}$$
$$H_A : \theta_i \neq \phi_i, \quad \text{for at least one } i, \tag{10.4}$$

so our null hypothesis is that the probability of a particular rating is the same for either restaurant.

10.3 A test statistic

Here, we use the test statistic

$$X^2 = \sum_{i=1}^{R} \sum_{j=1}^{C} \frac{(O_{i,j} - E_{i,j})^2}{E_{i,j}},$$

where

- $O_{i,j}$ is the observed count in row i, column j of the table,
- $E_{i,j}$ is the expected count in row i, column j of the table assuming H_0 is true, and is computed as

$$E_{i,j} = \frac{(\text{total in row } i) \times (\text{total in column } j)}{\text{grand total}}$$

- R is the number of rows in the table (2 in the example),
- C is the number of columns in the table (5 in the example).

- The test statistic cannot be negative. The smallest possible value is 0, when all the observations are exactly what we would expect to see, if H_0 were true.
- X^2 will get larger as the observed values $O_{i,j}$ differ more from the expected values $E_{i,j}$ under H_0; larger values of X^2 will give stronger evidence *against* H_0.

10.3.1 The formula for the expected counts

We now explain where the formula for $E_{i,j}$ comes from. As an example, consider $E_{1,2}$: the expected number of "very good" ratings for Akbar's. From the multinomial distribution that we defined previously ($Y \sim multinom(298; \theta_1, \theta_2, \ldots, \theta_5)$), this expected number would be

$$298 \times \theta_2,$$

so we will need an estimate of θ_2, the probability of a "very good" ratings for Akbar's. We estimate θ_2 using the following argument.

- Under H_0, the probability of a "very good" rating is the same in both rows of the table: we have $\theta_2 = \phi_2$.
- We observed $70 + 277 = 347$ "very good" ratings in total, out of a grand total of 1214 ratings
- An appropriate estimate of θ_2 would therefore be $\frac{374}{1214}$. (We would use the same estimate of ϕ_2, under H_0 which assumes $\theta_2 = \phi_2$).

Hence we have

$$E_{1,2} = 298 \times \frac{347}{1214} = \frac{(\text{total in row } 1) \times (\text{total in column } 2)}{\text{grand total}}.$$

10.4 Computing the test statistic for the observed data

We now compute all the expected values:

	Excellent	Very good	Average	Poor	Terrible
Akbar's	$\frac{298\times565}{1214}$	$\frac{298\times347}{1214}$	$\frac{298\times135}{1214}$	$\frac{298\times90}{1214}$	$\frac{298\times77}{1214}$
Aagrah	$\frac{916\times565}{1214}$	$\frac{916\times347}{1214}$	$\frac{916\times135}{1214}$	$\frac{916\times90}{1214}$	$\frac{916\times77}{1214}$

We can now compute our observed test statistic:

$$X^2_{obs} = \frac{(146-138.7)^2}{138.7} + \frac{(70-85.18)^2}{85.18} \tag{10.5}$$
$$+ \frac{(33-33.14)^2}{33.14} + \frac{(24-22.09)^2}{22.09} \tag{10.6}$$
$$+ \frac{(25-18.9)^2}{18.9} + \frac{(419-426.3)^2}{426.3} \tag{10.7}$$
$$+ \frac{(277-261.82)^2}{261.82} + \frac{(102-101.86)^2}{101.86} \tag{10.8}$$
$$+ \frac{(66-67.91)^2}{67.91} + \frac{(52-58.1)^2}{58.1} \tag{10.9}$$
$$= 6.92 \tag{10.10}$$

We'll also do this calculation in R. We first set up the table of observed value in R using the `matrix()` command, specifying 2 rows (`nrow = 2`), 5 columns (`ncol = 5`), and that the provided values (`c(146, 70, ..., 52)`) are specified in order along the **rows** (`byrow = TRUE`).

```
observed <- matrix(c(146, 70, 33, 24, 25,
                     419, 277, 102, 66, 52),
                   nrow = 2, ncol = 5, byrow = TRUE)
```

To check this has worked:

```
observed
```

```
##      [,1] [,2] [,3] [,4] [,5]
## [1,]  146   70   33   24   25
## [2,]  419  277  102   66   52
```

Now we'll compute a matrix of expected values, using the `outer()`, `rowSums` and `colSums()` commands. Don't worry about how these work (though you might guess), just note where we've used the matrix `observed`.

```
expected <- outer(rowSums(observed), colSums(observed),  "*") /
  sum(observed)
```

Then to compute the test statistic:

```
sum((observed - expected)^2 / expected)
```

```
## [1] 6.922
```

10.5 A simulation method

As in previous chapters, we will first try a simulation method to see how 'easy' it is to obtain a test statistic as large as the one we observed, if H_0 were true.

Assuming H_0 to be true, and using the estimated probabilities used in the calculation of the expected values $E_{i,j}$, we have

$$Y \sim multinom(298; \frac{565}{1214}, \frac{347}{1214}, \frac{135}{1214}, \frac{90}{1214}, \frac{77}{1214}), \tag{10.11}$$

$$Z \sim multinom(916; \frac{565}{1214}, \frac{347}{1214}, \frac{135}{1214}, \frac{90}{1214}, \frac{77}{1214}). \tag{10.12}$$

We now get R to simulate a random Y and Z, and assemble a new matrix of observed values, which we call `newObserved`.

```
Y <- rmultinom(1, 298, c(565, 347, 135, 90, 77)/1214)
Z <- rmultinom(1, 916, c(565, 347, 135, 90, 77)/1214)
newObserved <- matrix(c(Y, Z),
                      nrow = 2, ncol = 5, byrow = TRUE)
```

and to see what we got:

```
newObserved
```

```
##      [,1] [,2] [,3] [,4] [,5]
## [1,]  167   75   24   17   15
## [2,]  436  252  100   76   52
```

Now we compute the test statistic on this simulated data:

```
newExpected <- outer(rowSums(newObserved), colSums(newObserved),  "*") /
  sum(newObserved)
```

```
sum((newObserved - newExpected)^2 / newExpected)
```

```
## [1] 7.63
```

This has produced a larger test statistic: even though the probabilities of the different ratings were the same for the two restaurants, in the simulation, there was a bigger difference between the *observed* ratings for the two restaurants.

Now we repeat the simulation a large number of times.

```
chisquared <- rep(0, 100000)
for(i in 1:100000){
   Y <- rmultinom(1, 298, c(565, 347, 135, 90, 77)/1214)
   Z <- rmultinom(1, 916, c(565, 347, 135, 90, 77)/1214)
   newObserved <- matrix(c(Y, Z),
                         nrow = 2, ncol = 5, byrow = TRUE)
   newExpected <- outer(rowSums(newObserved),
                        colSums(newObserved),  "*") /
      sum(newObserved)
   chisquared[i] <- sum((newObserved - newExpected)^2 / newExpected)
}
sum(chisquared >= 6.92)
```

```
## [1] 14014
```

So about 14% of the time (14014/100000), we get test statistics larger than 6.92: if H_0 were true, there would be nothing 'surprising' about the observed differences in ratings between the two restaurants.

The 14% is our p-value: our probability that the test statistic would be as large as 6.92, if H_0 is true.

We'll draw a histogram of our test statistics to visualise this.

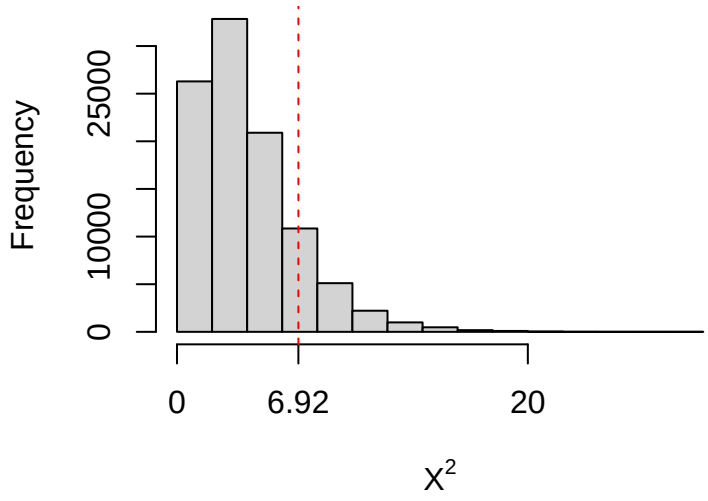


Figure 10.2: Histogram of randomly simulated test statistics, all simulated assuming H_0 is true. The larger values correspond to larger differences between the observed values, and the expected values under H_0. We are able to generate test statistics larger than the observed one (6.92) about 14% of the time.

10.6 An analytical method

Rather than using simulation, we can instead use an analytical approach, based on an approximate distribution of the test statistic under H_0.

If H_0 is true then, approximately

$$X^2 \sim \chi^2_\nu,$$

with $\nu =$ (number of rows -1) $\times$ (number of columns -1). We then compute the p-value as

$$P(X^2 \geq X^2_{obs}),$$

for $X^2 \sim \chi^2_\nu$.

As a rule-of-thumb, each expected count should be at least 5, otherwise the approximation may not be very good. If any expected count is less than 5, we could combine columns in the table, so that we have a smaller table, but more observations in (some of) the cells.

Informally, we can think of the degrees of freedom parameter ν as the number of 'pieces of information' we have to compare the two distributions.

- We have 10 observations in total: 2 rows $\times$ 5 columns.
- The row totals are fixed: the values in the last column are determined by the values in the first four columns. This reduces us to $2 \times (5-1)$ 'unconstrained' observations.

- We have had to estimate the probability of response in each column. Probabilities must sum to one, so we have estimated (number of columns -1) parameters here.
- The degrees of freedom is the number of 'unconstrained' observations, minus the the number of estimated parameters: $\nu =$ (number of rows)$\times$(number of columns -1) $-$ (number of columns$\}$ -1)=(number of rows$\}$ -1)$\times$(number of columns -1).

In our example, we have $\nu =$ (number of rows-1)$\times$ (number of columns -1) $= 4$, so we need to compute

$$P(X^2 \geq 6.92),$$

for $X^2 \sim \chi^2_4$. We obtain this from R with the `pchisq()` command:

```
1 - pchisq(6.92, 4)
```

```
## [1] 0.1402
```

Notice how close this is to the p-value obtained in our simulation. This is because the χ^2 approximation is good for this sample size. We illustrate this below.

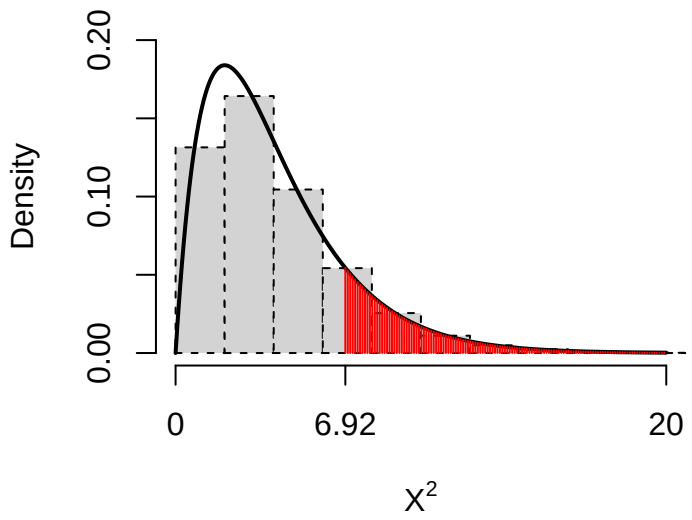


Figure 10.3: The χ^2_4 distribution, together with a histogram of our simulated test statistics. Notice the close agreement: the approximation of a χ^2_4 distribution for the test statistic is a good one. The p-value (0.14) is shown by the red shaded area.

In conclusion, we state that there is no evidence against H_0: no evidence to say that a customer is more likely to rate one restaurant higher than the other. This would suggest that the difference in rankings between the two restuarants (116 and 187) is not particularly meaningful.

10.6.1 χ^2 tests in R

We can conduct a χ^2 test in R using the command `chisq.test()`. Recall that we set up our data as a matrix in R:

```
observed
```

```
##      [,1] [,2] [,3] [,4] [,5]
## [1,]  146   70   33   24   25
## [2,]  419  277  102   66   52
```

We can then use the `chisq.test()` function:

```
chisq.test(observed)
```

```
##
##  Pearson's Chi-squared test
##
## data:  observed
## X-squared = 6.9, df = 4, p-value = 0.1
```

10.6.2 Row homogeneity and independence

The χ^2 test described above can be used in two different situations: testing for "row homogeneity" and testing for "independence". **The calculations used in each case are identical** and so for this module, you don't need to worry about the difference. In later modules, or in textbooks, you may come across these two terms, so we will explain what they mean. Two examples are given, which are included in the Tutorial Booklet, and so you will be able to see the solutions.

10.6.2.1 Row homogeneity

Here, we suppose that totals in each row are fixed in advance. For each row, the counts along the columns can be modelled with a multinomial distribution. By "row homogeneity", we mean that the multinomial probabilities are the same for every row. For example, suppose we want to compare voting preferences between voters who did and did not attend university. Suppose two random samples are drawn: 500 from each group. Participants named their favoured political party. Fictitious data are given below.

	Con	Lab	Lib	Other
University: yes	202	82	102	114
University: no	178	116	42	116

Here, the totals in each row are fixed at 500, and we can test whether the probabilities of voting for particular parties are the same in each row.

10.6.2.2 Independence

In this scenario, we take two 'measurements' per individual, and want to tests whether the measurements are related. For example, in a survey of 237 (Statistics) students from the University of Adelaide, smoking habits (recorded here as one of "Never", "Occasional", "Regular" or "Heavy") and exercise levels (recorded here as one of "Regular" or "some/none") were observed. A contingency table is given below, with smoking status in the rows, and exercise in the columns [^61].

Smoking status	exercise: regular	exercise: some/none
Never	87	102
Occasional	12	7
Regular	9	8
Heavy	7	4

Table 10.3: A contingency table showing numbers of students responses to a statement 'I am satisfied with the quality of this module', for two successive years. Percentages for each response within each each year are shown in brackets.

	Definitely agree	Mostly agree	Neither agree nor disagree	Mostly disagree	Definitely disagree
2017	15 (26.8%)	25 (44.6%)	9 (16.1%)	6 (10.7%)	1 (1.8%)
2018	12 (36.4%)	17 (51.5%)	3 (9.1%)	1 (3%)	0 (0%)

Here, we may wish to test whether smoking status is independent of exercise level, for example:

$$P(\text{regular exerciser and heavy smoker}) \tag{10.13}$$
$$= P(\text{regular exerciser}) \tag{10.14}$$
$$\times P(\text{heavy smoker}) \tag{10.15}$$

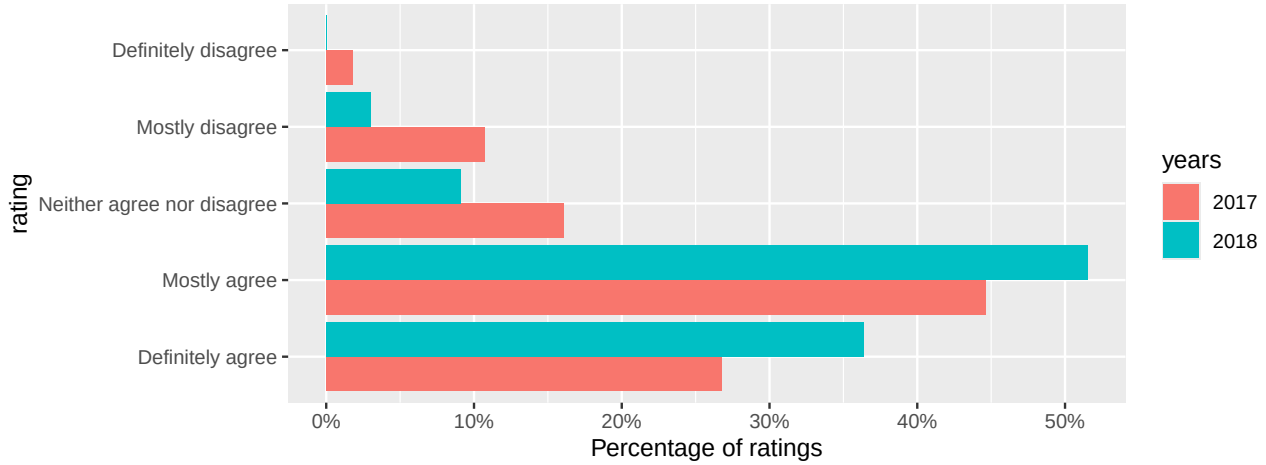
We can use **exactly the same** method as before: we calculate the expected values in exactly the same way, and compute the same X^2 test statistic, checking for cells with small expected counts as before. The reason this works is that we can express the null hypothesis of independence another way:

$$P(\text{regular exerciser} \mid \text{heavy smoker}) = P(\text{regular exerciser}),$$

so that the probability of being a regular exerciser would be the same for each row of the table: this is exactly what we would test for when testing for row homogeneity.

10.7 Exercise

Example 10.1 (Analysing student module questionnaire results)**.** The following data are from student module evaluations for this module, taken from two successive years. Regarding the statement, "I am satisfied with the quality of this module," the following responses were given. (56 students responded in 2017, and 33 students responded in 2018.)



The counts for the disagree responses are small, so we will combine those categories:

Year	Definitely agree	Mostly agree	Neither agree nor disagree / mostly disagree / definitely disagree
2017	15	25	16
2018	12	17	4

There appears to be some improvement in 2018, but is it statistically significant?

1. State the model and suitable hypotheses
2. Under a hypothesis of no difference between the years, give the expected number of "definitely agree" responses in 2018.
3. Given a χ^2 statistic of 3.34, draw a sketch to indicate what the p-value would represent, and give a conclusion for the test. Some of the R output below will help you.

```
pchisq(3.34, df = c(1, 2, 3, 4))
```

```
## [1] 0.9324 0.8118 0.6579 0.4974
```

Solution

1. We can model these data using multinomial distributions. Define

$$Y = (Y_1, Y_2, Y_3)$$

to be the numbers responding in each category in 2017, and

$$Z = (Z_1, Z_2, Z_3)$$

to be the numbers responding in each category in 2018 (so, for example, Y_2 would be the number of "mostly agree responses" in 2017, and Z_1 would be the number of "definitely agree" responses in 2018.) Before the responses are observed (but given the total number of responses in each year), we might suppose that

$$Y \sim multinom(56; \theta_1, \theta_2, \theta_3), \tag{10.16}$$
$$Z \sim multinom(33; \phi_1, \phi_2, \phi_3), \tag{10.17}$$

with, for example, θ_1 the probability of a "Definitely agree" response in 2017, and ϕ_2 the probability of a "Mostly agree" response in 2018.

For the null hypothesis, we suppose there was no difference between the two years: probabilities of any particular response were the same in each year. For the alternative hypothesis, at least some of the probabilities should be different. We write

$$H_0 : \theta_i = \phi_i, \quad i = 1, 2, 3 H_A \qquad : \theta_i \neq \phi_i, \quad \text{for at least one } i. \tag{10.18}$$

2. For the expected number of "definitely agree" responses in 2018, we compute

$$\frac{\text{total no. of 2018 responses} \times \text{total no. of "definitely agree responses"}}{\text{total number of responses}} = \frac{33 \times 27}{89}$$

3. Note that the X^2 statistic has 2 degrees of freedom, so the R output of interest is

The p-value is the probability of exceeding 3.34, i.e. $1 - 0.81 = 0.19$

The p-value is indicated by the shaded area: $1 - 0.81 = 0.19$. Hence, if H_0 were true, the differences we observed between the years would be 'unsurprising': there is no evidence against the null hypothesis of no change in student satisfaction.

Chapter 11

Index of definitions and examples

11.1 Definitions

11.2 Examples

- 8.1 Hypothesis testing: comparing binomial proportions. Can early release and tagging of prisoners affect the likelihood of reoffending?
- 7.1 Is quitting Facebook good for you?
- 6.1 One sample t-test: Energy drink company claim
- 9.1 Sample size and power calculation for a hypothesis test.
- 4.2 Standard error of the sample mean
- 4.6 Standard error of the sample proportion
- 4.3 Standard error of the sample variance
- 7.2 Testing a new diabetes treatment.
- 4.1 Unbiased estimators: sample mean and sample variance
- 4.5 Unbiased estimators: sample proportion